IDENTIFICATION OF RICE VARIETIES

Term End Milestone-1 (Project -1)

Proposal & Data Selection

Prashant Raghuwanshi (2223-1)

DSC630-T301 Predictive Analytics (2223-1)

Professor Catie Williams

DSC, Bellevue University

03/18/2022

Each phase of the process:

1. Business understanding
   a. Assess the Current Situation
      i. Inventory of resources
      ii. Requirements, assumptions and constraints
      iii. Risks and contingencies
      iv. Terminology
      v. Costs and benefits
   b. What are the Desired Outputs
   c. What Questions Are We Trying to Answer?
2. Data Understanding
   a. Initial Data Report
   b. Describe Data
   c. Initial Data Exploration
   d. Verify Data Quality
      i. Missing Data
      ii. Outliers
   e. Data Quality Report
3. Data Preparation
   a. Select Your Data
   b. Cleanse the Data
      i. Label Encoding
      ii. Drop Unnecessary Columns
      iii. Altering Datatypes
      iv. Dealing With Zeros
   c. Construct Required Data
   d. Integrate Data

4. Exploratory Data Analysis
5. Modelling
    a. Modelling Technique
    b. Modelling Assumptions
    c. Build Model
    d. Assess Model
6. Evaluation
7. Deployment

Template Source : https://www.sv-europe.com/crisp-dm-methodology/

Abstract:

This term-end project-1 aims to evaluate the Students should be able to identify a business problem to address through predictive analytics. The goal is to select appropriate models and model specifications and apply the respective methods to enhance data-driven decision-making related to the business problem. Students will identify the potential use of predictive analytics, formulate the problem, identify the right sources of data, analyze data and prescribe actions to improve not only the process of decision making but also the outcome of decisions. In this project, we are using datasets that contain, five different varieties of rice belonging to the same trademark were selected to carry out classification operations using morphological, shape, and color features. A total of 75 thousand rice grain images, including 15 thousand for each variety, were obtained. The images were pre-processed using MATLAB software and prepared for feature extraction. Using a combination of 12 morphological, 4 shape features, and 90 color features obtained from five different color spaces, a total of 106 features were extracted from the images. For classification, models were created with algorithms using machine learning techniques of k-nearest neighbor, decision tree, logistic regression, multilayer perceptron, random forest, and support vector machines. With these models, performance measurement values were obtained for feature sets of 12, 16, 90, and 106. Among the models, the success of the algorithms with the highest average classification accuracy was achieved 97.99% with random forest for morphological features. 98.04% were obtained with random forest for morphological and shape features. It was achieved with logistic regression as 99.25% for color features. Finally, 99.91% was obtained with multilayer perceptron for morphological, shape, and color features. When the results are examined, it is observed that with the addition of each new feature, the success of classification increases. Based on the performance measurement values obtained, it is possible to say that the study achieved success in classifying rice varieties.

1. Stage One - Determine Business Objectives and Assess the Situation

1.1 Assess the Current Situation

The modern Food Processing industry is importing grains (rice) from various international grains distributors. Their producted packed foods mostly depend on quality and varieties of imported raw gains. Even though placing the same type of rice variety order by food processing companies, most of the time Multiple gains distributors supplies the adulterated mix with the ordered rice varieties, and it results in causing the quality degradation and inconsistent taste of packed food and at last, it impacts the sales of processed food product in the competitive market place. At present food processing companies are using random sampling and manual grain monitoring techniques to make sure the procured variety of rice is good. However, this technique is not giving consistent results, since it depends on the individual human eye for identifying the quality of the grain from a single sample. Most of the time due to lack of expertise human eyes are not able to detect the ambiguities in a sample.

1.1.1. Inventory of resources

List the resources available to the project including:

- Personnel: Prashant Raghuwanshi
- Data: Rice_MSC_Dataset
- Computing resources: Personal computers
- Software: Jupyter Notebook(Python)

1.1.2. Requirements, assumptions and constraints -

Requirements : This Project is trying to make use of machine learning techniques to automatically identify the variety of the rice in the given rice sample. Here I am planning to feed the data for the rice to the ML model and the ML model will detect the rice sample and send a signal to the grain sampling machine’s sensors to pick out the ambiguous rice variety from the sample.

assumption: Here I am assuming the Preprocessing operations were applied to the rice images was applied successfully and made available data for feature extraction is not having any issue.

constraints : Major Constraints are related to used datasets and processed images, here the used datasets contain a total of 75 thousand rice grain images, including 15 thousand for each variety however due to the rapidly advancing Seed development process, we might not have all full collections of grain records under each gain varieties.

1.1.3.Risks and contingencies

- Risks include potential PII issues with respondents, which can largely be mitigated by anonymized respondent IDs. As long as a particular individual is never personally idenitified the risk of PII information leaking is mitigated.

- Potential for respondents to misrepresent their individual situations in their survey replies. Survey responses are to some extent unverifiable. For example, if a respondent indicates that the reason they remain unemployed is a long term medical issue, we can't verify that claim and that could potentially skew results.

- Lack of suitable candidates for a given application based on survey responses. This can be mitigated by the model indicating viability within the candidate pool.

1.1.4.Terminology

CRISP-DM The Cross-Industry Standard Process for Data-Mining – CRISP-DM is a model of a data mining process used to solve problems by experts. The model identifies the different stages in implementing a data mining project, as described bellow The model proposes the following steps: • Business Understanding – to understand the rules and business objectives of the company. • Understanding Data – to collect and describe data. • Data Preparation – to prepare data for import into the software. • Modeling – to select the modeling technique to be used. • Evaluation – Evaluate the process to see if the technique solves modeling and creating rules. • Deployment – to deploy the system and train its users

1.1.5.Costs and benefits

Costs: • The primary cost associated with this project is the time of the people working on it. • Computing resources for modeling • Data collection and processing computing costs

Benefits: • Social benefit of this model is helping the food processing companies to automatically detect the variety of rice in the provided sample of rice and helping the authority to stop low-cost mixing and adulteration practices of grain traders • Financial benefit to the company from the ability to maintain the brand quality of processed food which results in increasing the brand loyalty for consumers and its sales.

1.2 What Questions Are We Trying To Answer?

- Targeted Parameters:
- Can we identify survey respondent segments that are candidates for potential packaging as demographic data to sell to our customers?
- How can we determine how non technical professional time use data impacts ability of customers to find candidates for employment?
- How can we determine targeted parameters based on employer skill set requirements?
- How can we determine how technical professional time use data impacts ability of customers to find candidates for employment?
- Does a respondent's family status have an impact on their employment?
- Does the working status of a respondent's spouse impact their employment?
- What impact does non work time usage (Free time/Spending time with Children, etc.) have on respondent employment choices?

2. Stage Two - Data Understanding

The second stage of the CRISP-DM process requires you to acquire the data listed in the project resources. This initial collection includes data loading, if this is necessary for data understanding. For example, if you use a specific tool for data understanding, it makes perfect sense to load your data into this tool. If you acquire multiple data sources then you need to consider how and when you're going to integrate these.

A total of 75 thousand pieces of rice grain were obtained, including 15 thousand pieces of each variety of rice (Arborio, Basmati, Ipsala, Jasmine, Karacadag). Preprocessing operations were applied to the images and made available for feature extraction. A total of 106 features were inferred from the images; 12 morphological features and 4 shape features were obtained using morphological features and 90 color features were obtained from five different color spaces (RGB, HSV, Lab*, YCbCr, XYZ).

2.1 Initial Data Report

Initial data collection report - List the data sources acquired together with their locations, the methods used to acquire them and any problems encountered. Record problems you encountered and any resolutions achieved. This will help both with future replication of this project and with the execution of similar future projects.

```
# Import Libraries Required
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import numpy as np
import seaborn as sns
pd.set_option("display.max_columns", None)
from scipy.cluster.hierarchy import dendrogram
from sklearn.cluster import AgglomerativeClustering
from sklearn.preprocessing import StandardScaler, normalize
from sklearn.metrics import silhouette_score
import scipy.cluster.hierarchy as shc
from sklearn.decomposition import PCA
# supress warnings
import warnings
import seaborn as sns
from matplotlib.pyplot import xticks
from sklearn import preprocessing
from kmodes.kmodes import KModes
```

```
#Data source:
#Source Query location:
path =
'C:/Users/21313711/Documents/DSC680/Rice_MSC_Dataset/Rice_MSC_Dataset.
xlsx'
# reads the data from the file - denotes as CSV, it has no header,
sets column headers
df =  pd.read_excel(path)
```

```
df.shape
```

```
(75000, 107)
```

2.2 Describe Data

Attribute Information:

1.) Area: Returns the number of pixels within the boundaries of the rice grain.

2.) Perimeter: Calculates the circumference by calculating the distance between pixels around the boundaries of the rice grain.

3.) Major Axis Length: The longest line that can be drawn on the rice grain, i.e. the main axis distance, gives.

4.) Minor Axis Length: The shortest line that can be drawn on the rice grain, i.e. the small axis distance, gives.

5.) Eccentricity: It measures how round the ellipse, which has the same moments as the rice grain, is.

6.) Convex Area: Returns the pixel count of the smallest convex shell of the region formed by the rice grain.

7.) Extent: Returns the ratio of the regionformed by the rice grain to the bounding box pixels.

8.) Class: Cammeo and Osmancik rices

```
df.columns
```

```
Index(['AREA', 'PERIMETER', 'MAJOR_AXIS', 'MINOR_AXIS',
'ECCENTRICITY',
       'EQDIASQ', 'SOLIDITY', 'CONVEX_AREA', 'EXTENT', 'ASPECT_RATIO',
       ...
       'ALLdaub4L', 'ALLdaub4a', 'ALLdaub4b', 'ALLdaub4Y',
'ALLdaub4Cb',
       'ALLdaub4Cr', 'ALLdaub4XX', 'ALLdaub4YY', 'ALLdaub4ZZ',
'CLASS'],
      dtype='object', length=107)
```

```
df.shape
```

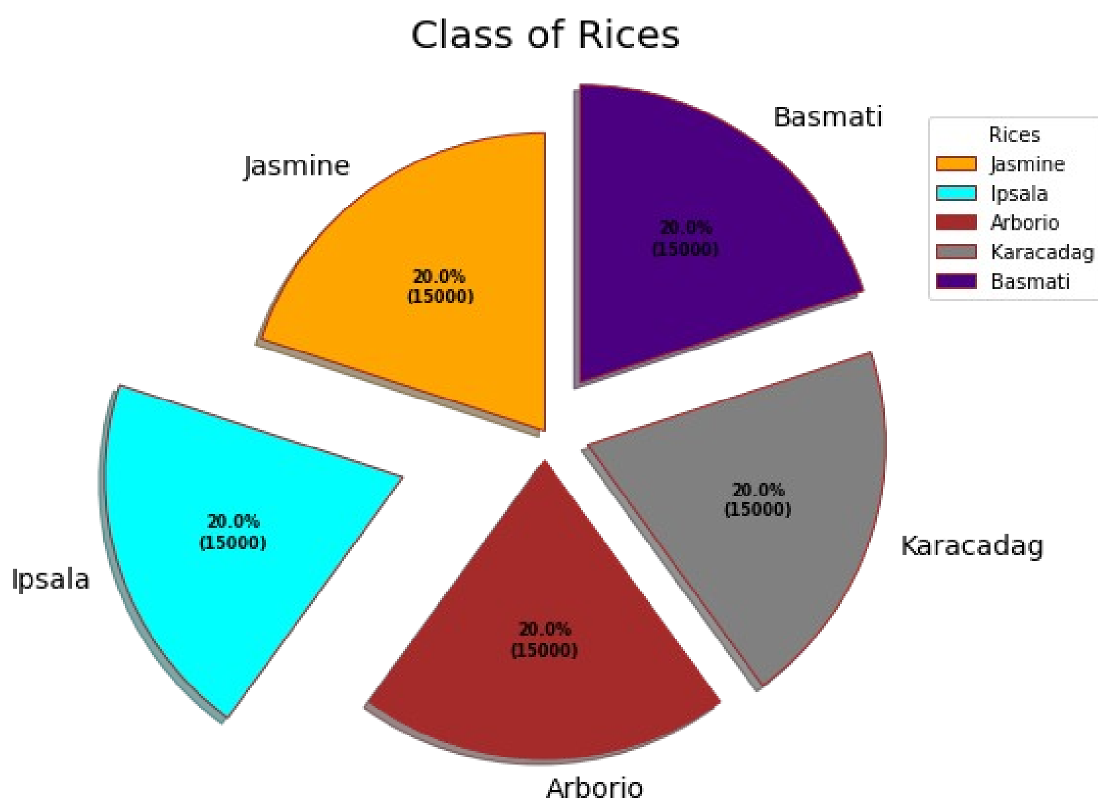
```
(75000, 107)
```

```
label = df["CLASS"].value_counts().index
value = df["CLASS"].value_counts().values
explode = (0.0, 0.5, 0.1, 0.15, 0.2)
colors = ( "orange", "cyan", "brown","grey", "indigo")
wp = { 'linewidth' : 1, 'edgecolor' : "brown" }

def func(pct, allvalues):
    absolute = int(pct / 100.*np.sum(allvalues))
    return "{:.1f}%\n({:d})".format(pct, absolute)
```

```
fig, ax = plt.subplots(figsize =(10, 7))
wedges, texts, autotexts = ax.pie(value,
                                  autopct = lambda pct: func(pct,
value),
                                  explode = explode,
                                  labels = label,
                                  shadow = True,
                                  colors = colors,
                                  startangle = 90,
                                  wedgeprops = wp,
                                  textprops = dict(color
="k",fontsize=14))

ax.legend(wedges, label,
          title ="Rices",
          loc ="center left",
          bbox_to_anchor =(1, 0, 0.8, 1.6))
plt.setp(autotexts, size = 8, weight ="bold")
ax.set_title("Class of Rices",fontsize=20)
plt.show()
```




```
df.describe()

                 AREA     PERIMETER    MAJOR_AXIS    MINOR_AXIS
ECCENTRICITY  \
count  75000.000000  75000.000000  75000.000000  75000.000000
```

```
75000.000000
mean    8379.197507    378.169453    161.805540     66.829335
0.886077
std     3119.209274     70.597008     36.461005     16.689269
0.071906
min     3929.000000    261.040000     96.968300     34.673000
0.627700
25%     6259.000000    316.431500    132.623500     49.650200
0.846100
50%     7345.000000    351.261000    149.343950     69.183900
0.885600
75%     8901.000000    444.986000    197.462025     75.814125
0.950800
max    21019.000000    593.698000    255.647200    113.441100
0.986800

            EQDIASQ      SOLIDITY   CONVEX_AREA        EXTENT
ASPECT_RATIO  \
count  75000.000000  75000.000000  75000.000000  75000.000000
75000.000000
mean     101.731251      0.975896   8584.862320      0.633226
2.597063
std       17.874070      0.007966   3189.298025      0.123795
0.968982
min       70.728800      0.877500   4032.000000      0.278800
1.284500
25%       89.270400      0.970900   6385.000000      0.561000
1.876100
50%       96.705500      0.976400   7532.000000      0.655800
2.153200
75%      106.457100      0.982200   9153.000000      0.727800
3.228700
max      163.591600      0.992100  21633.000000      0.901700
6.179500

          ROUNDNESS   COMPACTNESS  SHAPEFACTOR_1  SHAPEFACTOR_2  \
count  75000.000000  75000.000000   75000.000000   75000.000000
mean       0.732505      0.646079       0.020619       0.008407
std        0.138637      0.110787       0.005287       0.001903
min        0.392500      0.400600       0.011300       0.005100
25%        0.620600      0.551100       0.017000       0.006600
50%        0.775400      0.677100       0.018600       0.008700
75%        0.834500      0.725300       0.026200       0.009700
max        0.980000      0.879900       0.036900       0.013500

       SHAPEFACTOR_3  SHAPEFACTOR_4        meanRR        meanRG
meanRB  \
count   75000.000000   75000.000000  75000.000000  75000.000000
75000.000000
mean        0.429692       0.985509    216.398005    218.205782
```

```
227.918353
std         0.141146       0.007280     13.308330     13.646445
10.682523
min         0.160500       0.896200    153.800000    157.249900
160.158400
25%         0.303700       0.981600    206.605125    207.848625
220.927750
50%         0.458500       0.986400    215.118800    217.137550
228.801250
75%         0.526100       0.990700    225.016125    227.339300
236.171600
max         0.774300       0.999000    252.183700    252.323100
252.108500

           StdDevRR      StdDevRG      StdDevRB        skewRR
skewRG  \
count  75000.000000  75000.000000  75000.000000  75000.000000
75000.000000
mean      15.342766     15.449838     15.477779     -1.778549     -
1.938456
std        3.454178      3.562578      3.468618      0.948735
1.111904
min        6.817100      6.411700      6.417500     -6.938800     -
7.911800
25%       12.579400     12.741500     13.050675     -2.360500     -
2.522300
50%       15.542900     15.686150     15.539300     -1.608600     -
1.683350
75%       17.881000     18.032225     17.891800     -1.095000     -
1.136700
max       29.967400     30.765400     30.858000      0.917900
0.771900

             skewRB    kurtosisRR    kurtosisRG    kurtosisRB
entropyRR  \
count  75000.000000  75000.000000  75000.000000  75000.000000
7.500000e+04
mean      -2.360081     11.955533     12.944259     14.467290 -
4.426130e+09
std        0.950987      7.479528      9.302984      7.754649
2.239719e+09
min       -6.938200      1.841300      1.878100      1.885200 -
1.356042e+10
25%       -3.002300      6.481900      6.536450      8.425450 -
4.604666e+09
50%       -2.321100      9.727700     10.003650     13.436350 -
3.626555e+09
75%       -1.609700     15.068550     15.746325     18.253800 -
2.877646e+09
max        1.162400     75.201600     89.363100     71.980400 -
```

1.474496e+09

	entropyRG	entropyRB	meanH	meanS	meanV
count	7.500000e+04	7.500000e+04	75000.000000	75000.000000	75000.000000
mean	-4.509678e+09	-4.820370e+09	0.547699	0.060556	0.898100
std	2.268614e+09	1.994516e+09	0.185565	0.036708	0.043447
min	-1.383477e+10	-1.317801e+10	0.034100	0.001400	0.628100
25%	-4.870068e+09	-5.477346e+09	0.526275	0.027200	0.868600
50%	-3.674473e+09	-4.153602e+09	0.644800	0.054250	0.903500
75%	-2.890273e+09	-3.415980e+09	0.664600	0.091700	0.932300
max	-1.550952e+09	-1.605495e+09	0.817100	0.241700	0.990600

	StdDevH	StdDevS	StdDevV	skewH	skewS
count	75000.000000	75000.000000	75000.000000	75000.000000	75000.000000
mean	0.064218	0.019138	0.060251	-4.797680	0.019438
std	0.061397	0.010459	0.013644	7.194686	1.043360
min	0.002100	0.003000	0.025100	-70.866500	-2.713400
25%	0.022200	0.010900	0.050500	-8.312225	-0.743300
50%	0.041700	0.015800	0.060700	-3.579100	-0.277450
75%	0.087425	0.025200	0.069900	-0.169475	0.810825
max	0.410300	0.093900	0.118400	25.021800	6.927700

	skewV	kurtosisH	kurtosisS	kurtosisV	entropyH
count	75000.000000	75000.000000	75000.000000	75000.000000	75000.000000
mean	-2.489326	131.841205	4.601397	15.402438	2309.978182
std	1.051619	261.985126	2.819120	9.049990	602.876603
min	-7.910400	1.009300	1.275100	1.885000	262.201600

```
25%       -3.149800      8.462025      2.874975      8.655600
1944.970550
50%       -2.428750     41.961300      4.007750     13.850600
2338.881600
75%       -1.659675    146.237675      5.489300     19.087700
2713.329300
max        0.777400   5504.557100     76.759400     89.212900
4868.357900

           entropyS      entropyV         meanL         meanA
meanB  \
count  75000.000000  75000.000000  75000.000000  75000.000000
75000.000000
mean     184.779482   1246.241590    222.215488    128.759108
122.920756
std      159.467177    468.043347     11.801014      2.352168
4.873599
min        1.619600    137.279200    164.704200    118.268500
107.303700
25%       52.218900    910.651125    213.312575    127.767675
118.926575
50%      123.785950   1194.487450    221.581450    128.838500
122.847600
75%      291.221550   1518.092750    230.164650    130.348625
126.154775
max      975.833900   4814.083500    252.505700    134.923800
140.567600

           StdDevL       StdDevA       StdDevB         skewL
skewA  \
count  75000.000000  75000.000000  75000.000000  75000.000000
75000.000000
mean      14.127218      0.939870      2.215095     -2.083832
0.114552
std        3.247309      0.395952      1.327015      1.072309
0.914165
min        5.840700      0.106100      0.000000     -7.911300     -
8.295200
25%       11.702575      0.629200      1.177900     -2.675600     -
0.467425
50%       14.405750      0.818300      1.674700     -1.863400
0.039600
75%       16.445050      1.220200      3.043725     -1.296075
0.756100
max       27.440700      3.258200     10.821100      0.671300
9.208500

             skewB     kurtosisL     kurtosisA     kurtosisB
entropyL  \
count  74994.000000  75000.000000  75000.000000  74994.000000
```

```
7.500000e+04
mean       0.529481     13.850684      4.282437      4.730635 -
4.654252e+09
std        0.997105      9.346636      2.361591      2.966087
2.248731e+09
min       -3.168200      1.918700      1.000000      0.999900 -
1.389150e+10
25%       -0.049300      7.333000      2.800700      2.905700 -
4.995975e+09
50%        0.581600     11.042450      3.702300      3.927800 -
3.842979e+09
75%        1.081900     16.932775      5.058150      5.611475 -
3.060054e+09
max        8.540500     89.816000     85.788600     73.882000 -
1.690136e+09

           entropyA      entropyB         meanY        meanCb
meanCr  \
count  7.500000e+04  7.500000e+04  75000.000000  75000.000000
75000.000000
mean  -1.342112e+09 -1.248950e+09    203.886710    132.483100
126.403359
std    4.750043e+08  5.704776e+08     10.866061      4.320342
2.350780
min   -3.267201e+09 -3.472658e+09    150.474500    116.642000
114.724000
25%   -1.404111e+09 -1.305433e+09    195.709775    129.684175
126.104525
50%   -1.196488e+09 -1.037515e+09    203.406400    132.584850
127.043700
75%   -1.035494e+09 -8.441662e+08    211.211900    136.026325
127.467425
max   -6.293092e+08 -5.472335e+08    232.550000    146.855400
133.076200

            StdDevY      StdDevCb      StdDevCr         skewY
skewCb  \
count  75000.000000  75000.000000  75000.000000  75000.000000
74997.000000
mean      13.163794      1.969016      0.754401     -1.955500     -
0.471430
std        2.979435      1.154209      0.407306      1.014331
1.226905
min        5.589200      0.000000      0.000000     -7.531800     -
7.501200
25%       10.892900      1.078000      0.468500     -2.521600     -
1.080800
50%       13.379150      1.511050      0.687700     -1.750650     -
0.586200
75%       15.288100      2.700025      0.923100     -1.210200
```

```
0.100100
max       25.460100      9.474100      4.167300      0.866100
120.657500

             skewCr     kurtosisY    kurtosisCb    kurtosisCr
entropyY  \
count  74998.000000  75000.000000  74997.000000  74998.000000
7.500000e+04
mean       1.734378     12.894485      5.128738     67.690019 -
3.856125e+09
std        7.868774      8.536443     58.358394    560.177674
1.862941e+09
min       -9.581300      1.871700      1.000000      0.999900 -
1.152219e+10
25%        0.140725      6.914600      2.847700      2.638800 -
4.134462e+09
50%        0.606200     10.370950      3.959400      3.572000 -
3.187489e+09
75%        1.343700     15.792500      5.734300      5.356475 -
2.539634e+09
max      116.318200     83.392400  14559.161100  13529.946300 -
1.380124e+09

          entropyCb     entropyCr        meanXX        meanYY
meanZZ  \
count  7.500000e+04  7.500000e+04  75000.000000  75000.000000
75000.000000
mean  -1.414347e+09 -1.300141e+09      0.684125      0.714363
0.842659
std    4.388745e+08  5.041548e+08      0.083806      0.094242
0.086255
min   -3.219491e+09 -3.294497e+09      0.319800      0.338400
0.384200
25%   -1.578173e+09 -1.336365e+09      0.620200      0.642100
0.782700
50%   -1.288827e+09 -1.134926e+09      0.680400      0.706700
0.848700
75%   -1.126542e+09 -9.659853e+08      0.740700      0.776000
0.909600
max   -6.653100e+08 -5.999740e+08      0.927600      0.977000
1.061300

           StdDevXX      StdDevYY      StdDevZZ       skewXX
skewYY  \
count  75000.000000  75000.000000  75000.000000  75000.00000
75000.000000
mean       0.097866      0.103307      0.116325     -1.15768     -
1.131194
std        0.021346      0.023054      0.025187      0.82280
0.900195
```

min	0.048300	0.049700	0.053300	-5.72170	-6.170300
25%	0.079700	0.083600	0.096700	-1.60470	-1.590575
50%	0.097200	0.102000	0.116600	-0.99310	-0.932550
75%	0.113400	0.120500	0.133100	-0.58500	-0.510000
max	0.194900	0.211900	0.226200	1.68760	1.633800

	skewZZ	kurtosisXX	kurtosisYY	kurtosisZZ	entropyXX \
count	75000.000000	75000.000000	75000.000000	75000.000000	75000.000000
mean	-1.504015	8.279693	8.306708	9.274027	2588.904041
std	0.829523	5.298437	5.769769	5.151446	743.477905
min	-5.953800	1.693700	1.698000	1.691000	764.255100
25%	-2.028425	4.476275	4.255000	5.220350	2089.496050
50%	-1.416100	6.707100	6.432400	8.568450	2395.386350
75%	-0.883975	10.304150	10.273325	11.667725	2805.124300
max	1.864400	53.962900	58.776300	49.425100	6322.974600

	entropyYY	entropyZZ	ALLdaub4RR	ALLdaub4RG	ALLdaub4RB \
count	75000.000000	75000.000000	75000.000000	75000.000000	75000.000000
mean	2367.177736	1489.687843	108.178754	109.082089	113.936257
std	596.172238	828.002085	6.657980	6.827630	5.343310
min	343.706900	-2074.590800	76.843600	78.572300	80.027800
25%	1975.926700	963.928125	103.278675	103.900075	110.438275
50%	2268.014150	1481.069350	107.534300	108.545600	114.379400
75%	2574.653375	1924.440875	112.486075	113.654875	118.063450
max	5835.086900	5615.509800	126.105600	126.169700	126.067200

	ALLdaub4H	ALLdaub4S	ALLdaub4V	ALLdaub4L

```
ALLdaub4a  \
count  75000.000000  75000.000000  75000.000000  75000.000000
75000.000000
mean       0.273845      0.030271      0.448960    111.088252
64.379443
std        0.092785      0.018347      0.021736      5.904854
1.175616
min        0.017200      0.000700      0.313900     82.300600
59.137900
25%        0.263100      0.013600      0.434200    106.632900
63.883800
50%        0.322400      0.027100      0.451600    110.770700
64.419350
75%        0.332300      0.045800      0.466100    115.065075
65.174200
max        0.408700      0.120800      0.495100    126.265100
67.459000

          ALLdaub4b     ALLdaub4Y    ALLdaub4Cb    ALLdaub4Cr
ALLdaub4XX  \
count  75000.000000  75000.000000  75000.000000  75000.000000
75000.000000
mean      61.461457    101.925425     66.240541     63.202088
0.341944
std        2.435635      5.436861      2.159109      1.174976
0.041921
min       53.653800     75.191800     58.323800     57.363400
0.159700
25%       59.465575     97.834400     64.842000     63.052800
0.309900
50%       61.424400    101.683700     66.291600     63.522050
0.340100
75%       63.076825    105.592450     68.011800     63.734000
0.370300
max       70.284000    116.287300     73.424700     66.539100
0.463900

         ALLdaub4YY    ALLdaub4ZZ
count  75000.000000  75000.000000
mean       0.357058      0.421176
std        0.047139      0.043137
min        0.169000      0.191800
25%        0.320900      0.391200
50%        0.353300      0.424200
75%        0.387900      0.454700
max        0.488600      0.530200

df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 75000 entries, 0 to 74999
Columns: 107 entries, AREA to CLASS
dtypes: float64(95), int64(11), object(1)
memory usage: 61.2+ MB
```

```
df.head(5)
```

```
   AREA  PERIMETER  MAJOR_AXIS  MINOR_AXIS  ECCENTRICITY   EQDIASQ
SOLIDITY  \
0  7805    437.915    209.8215     48.0221        0.9735   99.6877
0.9775
1  7503    340.757    138.3361     69.8417        0.8632   97.7400
0.9660
2  5124    314.617    141.9803     46.5784        0.9447   80.7718
0.9721
3  7990    437.085    201.4386     51.2245        0.9671  100.8622
0.9659
4  7433    342.893    140.3350     68.3927        0.8732   97.2830
0.9831

   CONVEX_AREA  EXTENT  ASPECT_RATIO  ROUNDNESS  COMPACTNESS
SHAPEFACTOR_1  \
0         7985  0.3547        4.3693     0.5114       0.4751
0.0269
1         7767  0.6637        1.9807     0.8120       0.7065
0.0184
2         5271  0.4760        3.0482     0.6505       0.5689
0.0277
3         8272  0.6274        3.9325     0.5256       0.5007
0.0252
4         7561  0.6006        2.0519     0.7944       0.6932
0.0189

   SHAPEFACTOR_2  SHAPEFACTOR_3  SHAPEFACTOR_4    meanRR    meanRG
meanRB  \
0         0.0062         0.2257         0.9863  222.9805  223.9872
241.4758
1         0.0093         0.4992         0.9888  206.0380  206.2412
213.3809
2         0.0091         0.3236         0.9865  201.8228  217.6475
235.0057
3         0.0064         0.2507         0.9859  228.8978  229.7151
244.6294
4         0.0092         0.4806         0.9860  210.4471  210.2988
220.8827

   StdDevRR  StdDevRG  StdDevRB  skewRR  skewRG  skewRB  kurtosisRR  \
0   16.2950   16.4354   13.6272 -0.7986 -0.8407 -3.7377      7.1706
1   18.3863   18.5343   18.9969 -0.7536 -0.7372 -0.8217      4.4926
```

```
2   13.7392   15.3239   16.0249 -2.2606 -2.6764 -2.8690     12.4329
3   18.3915   18.9141   16.7398 -1.5281 -1.4967 -3.5705      8.1541
4   19.6370   19.6635   20.0206 -1.0165 -1.0176 -1.1514      4.6467

   kurtosisRG  kurtosisRB   entropyRR   entropyRG   entropyRB   meanH
meanS  \
0      7.1197     23.5566 -4225152256 -4267165440 -5014238208  0.6532
0.0780
1      4.4713      4.4922 -3428633856 -3436527616 -3701304832  0.6542
0.0354
2     13.0172     13.4749 -2228499200 -2629524480 -3108116224  0.5875
0.1409
3      7.8216     17.9112 -4586280960 -4623998464 -5290343936  0.6155
0.0658
4      4.6610      5.1408 -3560909568 -3555586304 -3955763968  0.6677
0.0485

    meanV  StdDevH  StdDevS  StdDevV    skewH   skewS   skewV
kurtosisH  \
0  0.9470   0.0516   0.0358   0.0534  -6.2345 -0.3859 -3.7366
79.7876
1  0.8368   0.0535   0.0143   0.0745  -9.7857  0.8650 -0.8208
119.7635
2  0.9216   0.0131   0.0183   0.0628   1.4178 -0.8871 -2.8691
11.9584
3  0.9595   0.1407   0.0385   0.0655  -3.5418 -0.1676 -3.5726
15.2393
4  0.8662   0.0210   0.0107   0.0785 -21.9823  0.3104 -1.1504
693.8591

   kurtosisS  kurtosisV   entropyH  entropyS   entropyV     meanL
meanA  \
0     2.1248    23.5552  2802.0408  265.6183   701.6497  228.0955
130.1179
1     3.4847     4.4913  2696.2507   68.2768  1764.4907  211.4882
128.7362
2     5.5929    13.4749  1879.6370  400.6435   656.2015  219.5361
125.2797
3     1.7312    17.9282  2743.6218  219.5903   514.2898  233.0965
129.8081
4     4.1654     5.1382  2672.0652  108.4178  1481.3774  215.4514
129.4134

      meanB  StdDevL  StdDevA  StdDevB   skewL   skewA   skewB
kurtosisL  \
0  119.2036  14.5928   1.6133   4.3298 -1.1189  0.0608  0.3568
8.6096
1  124.4154  17.0760   0.7826   1.7041 -0.8311  0.9045 -0.5643
4.8704
2  117.4914  13.8545   1.0867   1.7036 -2.7675  1.2988  0.4409
```

```
13.8206
3  120.5208  16.8280   1.7279   4.8413 -1.7808  0.2575  0.0997
9.3138
4  122.7094  18.1186   0.6736   1.1177 -1.1075  0.3288  0.3235
5.0818

   kurtosisA  kurtosisB    entropyL    entropyA    entropyB     meanY
\
0     2.1559     2.1246 -4432724480 -1286927360 -1062276608  209.8105

1     3.7122     2.4539 -3623472640 -1208159104 -1120711552  193.7512

2     7.0239     5.1960 -2676372224  -777033216  -674463936  200.5582

3     2.0849     1.7092 -4764499456 -1310564736 -1114603904  214.5264

4     3.0013     5.0546 -3740926720 -1210803200 -1076777088  197.6947


     meanCb    meanCr  StdDevY  StdDevCb  StdDevCr   skewY  skewCb
skewCr  \
0  135.7502  126.1657  13.5906    3.7538    0.9819 -1.0936 -0.4188
0.1943
1  131.0736  127.1655  15.8972    1.3851    0.3717 -0.7533  0.7497
1.8000
2  137.9440  119.6659  12.7435    1.5291    1.6526 -2.6443 -0.6828
1.5534
3  134.6532  126.4443  15.6660    4.2306    0.9329 -1.7634 -0.1586
0.1566
4  132.4079  127.0735  16.9122    1.0164    0.3185 -1.0310 -0.0866
1.3699

   kurtosisY  kurtosisCb  kurtosisCr    entropyY   entropyCb
entropyCr  \
0     8.3887      2.1507      3.3297 -3693353216 -1414067840 -
1202124032
1     4.4987      2.9312      4.2389 -2992325376 -1257189760 -
1175834752
2    13.1546      6.0969      5.4741 -2196389376  -960911488  -
702334144
3     9.0651      1.6830      3.2673 -3974825984 -1422344704 -
1236613248
4     4.7069      4.8858     10.0536 -3100359680 -1273499776 -
1163001088

   meanXX  meanYY  meanZZ  StdDevXX  StdDevYY  StdDevZZ  skewXX
skewYY  \
0  0.7355  0.7594  0.9475    0.1067    0.1166    0.1080 -0.4876 -
0.2667
```

```
1  0.6035  0.6294  0.7289    0.1152    0.1207    0.1386 -0.3131 -
0.2934
2  0.6472  0.6897  0.8902    0.0883    0.0953    0.1203 -1.9035 -
1.9296
3  0.7762  0.8041  0.9809    0.1218    0.1330    0.1301 -1.2292 -
0.9592
4  0.6365  0.6605  0.7853    0.1240    0.1293    0.1497 -0.6071 -
0.5923

   skewZZ  kurtosisXX  kurtosisYY  kurtosisZZ  entropyXX  entropyYY  \
0 -2.4022      5.4413      4.6547     13.5992  2388.4263  2222.7087
1 -0.3768      2.9232      2.9286      2.8833  2567.9663  2531.9880
2 -2.2574      9.4334      9.4209      9.8885  1790.9473  1715.9330
3 -2.8063      6.0941      5.1355     12.2990  2162.7341  1909.6716
4 -0.6824      3.0376      3.0058      3.2231  2493.1421  2432.7375

   entropyZZ  ALLdaub4RR  ALLdaub4RG  ALLdaub4RB  ALLdaub4H  ALLdaub4S
\
0   512.8892    111.4315    111.9330    120.6838     0.3266     0.0390

1  2189.7100    102.9773    103.0778    106.6464     0.3270     0.0177

2   757.2745    100.8594    108.7688    117.4546     0.2938     0.0705

3   -63.9162    114.4421    114.8475    122.3142     0.3076     0.0329

4  1815.4894    105.2504    105.1742    110.4669     0.3338     0.0242


   ALLdaub4V  ALLdaub4L  ALLdaub4a  ALLdaub4b  ALLdaub4Y
ALLdaub4Cb  \
0     0.4733   113.9924    65.0610    59.5989   104.8552     67.8779

1     0.4182   105.7055    64.3685    62.2084    96.8375     65.5371

2     0.4606   109.7155    62.6423    58.7439   100.2352     68.9753

3     0.4797   116.5405    64.9069    60.2562   107.2560     67.3298

4     0.4332   107.7502    64.7071    61.3549    98.8704     66.2048


   ALLdaub4Cr  ALLdaub4XX  ALLdaub4YY  ALLdaub4ZZ    CLASS
0     63.0828      0.3673      0.3793      0.4733  Basmati
1     63.5832      0.3014      0.3144      0.3641  Arborio
2     59.8342      0.3233      0.3445      0.4448  Jasmine
3     63.2237      0.3880      0.4020      0.4904  Basmati
4     63.5378      0.3184      0.3303      0.3928  Arborio
```

2.3 Verify Data Quality

Examine the quality of the data, addressing questions such as:

- Is the data complete (does it cover all the cases required)?
- Is it correct, or does it contain errors and, if there are errors, how common are they?
- Are there missing values in the data? If so, how are they represented, where do they occur, and how common are they?

2.3.1. Missing Data

In addition to incorrect datatypes, another common problem when dealing with real-world data is missing values. These can arise for many reasons and have to be either filled in or removed before we train a machine learning model. First, let's get a sense of how many missing values are in each column

While we always want to be careful about removing information, if a column has a high percentage of missing values, then it probably will not be useful to our model. The threshold for removing columns should depend on the problem

```
# checking null value
df.isnull().sum()
```

```
AREA            0
PERIMETER       0
MAJOR_AXIS      0
MINOR_AXIS      0
ECCENTRICITY    0
               ..
ALLdaub4Cr      0
ALLdaub4XX      0
ALLdaub4YY      0
ALLdaub4ZZ      0
CLASS           0
Length: 107, dtype: int64
```

```
# This function take a dataframe
# as a parameter and returning list
# of column names whose contents
# are duplicates.
def getDuplicateColumns(df):

    # Create an empty set
    duplicateColumnNames = set()

    # Iterate through all the columns
    # of dataframe
    for x in range(df.shape[1]):

        # Take column at xth index.
```

```python
        col = df.iloc[:, x]

        # Iterate through all the columns in
        # DataFrame from (x + 1)th index to
        # last index
        for y in range(x + 1, df.shape[1]):

            # Take column at yth index.
            otherCol = df.iloc[:, y]

            # Check if two columns at x & y
            # index are equal or not,
            # if equal then adding
            # to the set
            if col.equals(otherCol):
                duplicateColumnNames.add(df.columns.values[y])

    # Return list of unique column names
    # whose contents are duplicates.
    return list(duplicateColumnNames)
```

```python
drop_columns2= getDuplicateColumns(df)
# list duplicate data columns
drop_columns2
```

```
[]
```

```python
# drop columns from df whose of values are duplicate
df.drop(drop_columns2, axis=1, inplace=True)
```

```python
# filling out missing values
df = df.fillna(method="ffill")
df = df.fillna(method="bfill")
```

3. Stage Three - Data Preperation

This is the stage of the project where you decide on the data that you're going to use for analysis. The criteria you might use to make this decision include the relevance of the data to your data mining goals, the quality of the data, and also technical constraints such as limits on data volume or data types. Note that data selection covers selection of attributes (columns) as well as selection of records (rows) in a table.

3.1 Label Encoding

```python
df.dtypes
```

```
AREA                int64
PERIMETER         float64
MAJOR_AXIS        float64
```

```
MINOR_AXIS      float64
ECCENTRICITY    float64
                 ...
ALLdaub4Cr      float64
ALLdaub4XX      float64
ALLdaub4YY      float64
ALLdaub4ZZ      float64
CLASS            object
Length: 107, dtype: object
```

- All fields are already label encoded. No need to change data types.
- May potentially update to One Hot Encoding.

3.2.2 Drop Unnecessary Columns

Sometimes we may not need certain columns. We can drop to keep only relevent data

3.2 Dealing With Zeros

Replacing all the zeros from cols. **Note** You may not want to do this - add / remove as required

- Zero values were previously addressed using mean value imputation.

3.3 Construct Required Data

This task includes constructive data preparation operations such as the production of derived attributes or entire new records, or transformed values for existing attributes.

Derived attributes - These are new attributes that are constructed from one or more existing attributes in the same record, for example you might use the variables of length and width to calculate a new variable of area.

Generated records - Here you describe the creation of any completely new records. For example you might need to create records for customers who made no purchase during the past year. There was no reason to have such records in the raw data, but for modelling purposes it might make sense to explicitly represent the fact that particular customers made zero purchases.

- Do not have derivative records. No construction required.

4. Stage Four - Exploratory Data Analysis

```
df.head()
```

```
   AREA  PERIMETER  MAJOR_AXIS  MINOR_AXIS  ECCENTRICITY   EQDIASQ
SOLIDITY  \
0  7805    437.915    209.8215     48.0221        0.9735   99.6877
0.9775
1  7503    340.757    138.3361     69.8417        0.8632   97.7400
```

```
0.9660
2  5124    314.617    141.9803     46.5784        0.9447   80.7718
0.9721
3  7990    437.085    201.4386     51.2245        0.9671  100.8622
0.9659
4  7433    342.893    140.3350     68.3927        0.8732   97.2830
0.9831

   CONVEX_AREA  EXTENT  ASPECT_RATIO  ROUNDNESS  COMPACTNESS
SHAPEFACTOR_1  \
0         7985  0.3547        4.3693     0.5114       0.4751
0.0269
1         7767  0.6637        1.9807     0.8120       0.7065
0.0184
2         5271  0.4760        3.0482     0.6505       0.5689
0.0277
3         8272  0.6274        3.9325     0.5256       0.5007
0.0252
4         7561  0.6006        2.0519     0.7944       0.6932
0.0189

   SHAPEFACTOR_2  SHAPEFACTOR_3  SHAPEFACTOR_4    meanRR    meanRG
meanRB  \
0         0.0062         0.2257         0.9863  222.9805  223.9872
241.4758
1         0.0093         0.4992         0.9888  206.0380  206.2412
213.3809
2         0.0091         0.3236         0.9865  201.8228  217.6475
235.0057
3         0.0064         0.2507         0.9859  228.8978  229.7151
244.6294
4         0.0092         0.4806         0.9860  210.4471  210.2988
220.8827

   StdDevRR  StdDevRG  StdDevRB  skewRR  skewRG  skewRB  kurtosisRR  \
0   16.2950   16.4354   13.6272 -0.7986 -0.8407 -3.7377      7.1706
1   18.3863   18.5343   18.9969 -0.7536 -0.7372 -0.8217      4.4926
2   13.7392   15.3239   16.0249 -2.2606 -2.6764 -2.8690     12.4329
3   18.3915   18.9141   16.7398 -1.5281 -1.4967 -3.5705      8.1541
4   19.6370   19.6635   20.0206 -1.0165 -1.0176 -1.1514      4.6467

   kurtosisRG  kurtosisRB   entropyRR   entropyRG   entropyRB   meanH
meanS  \
0      7.1197     23.5566 -4225152256 -4267165440 -5014238208  0.6532
0.0780
1      4.4713      4.4922 -3428633856 -3436527616 -3701304832  0.6542
0.0354
2     13.0172     13.4749 -2228499200 -2629524480 -3108116224  0.5875
0.1409
3      7.8216     17.9112 -4586280960 -4623998464 -5290343936  0.6155
```

```
0.0658
4      4.6610      5.1408 -3560909568 -3555586304 -3955763968  0.6677
0.0485

    meanV  StdDevH  StdDevS  StdDevV    skewH   skewS   skewV
kurtosisH  \
0  0.9470   0.0516   0.0358   0.0534  -6.2345 -0.3859 -3.7366
79.7876
1  0.8368   0.0535   0.0143   0.0745  -9.7857  0.8650 -0.8208
119.7635
2  0.9216   0.0131   0.0183   0.0628   1.4178 -0.8871 -2.8691
11.9584
3  0.9595   0.1407   0.0385   0.0655  -3.5418 -0.1676 -3.5726
15.2393
4  0.8662   0.0210   0.0107   0.0785 -21.9823  0.3104 -1.1504
693.8591

   kurtosisS  kurtosisV   entropyH  entropyS   entropyV     meanL
meanA  \
0     2.1248    23.5552  2802.0408  265.6183   701.6497  228.0955
130.1179
1     3.4847     4.4913  2696.2507   68.2768  1764.4907  211.4882
128.7362
2     5.5929    13.4749  1879.6370  400.6435   656.2015  219.5361
125.2797
3     1.7312    17.9282  2743.6218  219.5903   514.2898  233.0965
129.8081
4     4.1654     5.1382  2672.0652  108.4178  1481.3774  215.4514
129.4134

      meanB  StdDevL  StdDevA  StdDevB   skewL   skewA   skewB
kurtosisL  \
0  119.2036  14.5928   1.6133   4.3298 -1.1189  0.0608  0.3568
8.6096
1  124.4154  17.0760   0.7826   1.7041 -0.8311  0.9045 -0.5643
4.8704
2  117.4914  13.8545   1.0867   1.7036 -2.7675  1.2988  0.4409
13.8206
3  120.5208  16.8280   1.7279   4.8413 -1.7808  0.2575  0.0997
9.3138
4  122.7094  18.1186   0.6736   1.1177 -1.1075  0.3288  0.3235
5.0818

   kurtosisA  kurtosisB    entropyL    entropyA    entropyB     meanY
\
0     2.1559     2.1246 -4432724480 -1286927360 -1062276608  209.8105

1     3.7122     2.4539 -3623472640 -1208159104 -1120711552  193.7512

2     7.0239     5.1960 -2676372224  -777033216  -674463936  200.5582
```

```
3     2.0849     1.7092 -4764499456 -1310564736 -1114603904  214.5264

4     3.0013     5.0546 -3740926720 -1210803200 -1076777088  197.6947


     meanCb    meanCr  StdDevY  StdDevCb  StdDevCr   skewY  skewCb  
skewCr  \
0  135.7502  126.1657  13.5906    3.7538    0.9819 -1.0936 -0.4188   
0.1943
1  131.0736  127.1655  15.8972    1.3851    0.3717 -0.7533  0.7497   
1.8000
2  137.9440  119.6659  12.7435    1.5291    1.6526 -2.6443 -0.6828   
1.5534
3  134.6532  126.4443  15.6660    4.2306    0.9329 -1.7634 -0.1586   
0.1566
4  132.4079  127.0735  16.9122    1.0164    0.3185 -1.0310 -0.0866   
1.3699

   kurtosisY  kurtosisCb  kurtosisCr    entropyY   entropyCb   
entropyCr  \
0     8.3887      2.1507      3.3297 -3693353216 -1414067840 -
1202124032
1     4.4987      2.9312      4.2389 -2992325376 -1257189760 -
1175834752
2    13.1546      6.0969      5.4741 -2196389376  -960911488  -
702334144
3     9.0651      1.6830      3.2673 -3974825984 -1422344704 -
1236613248
4     4.7069      4.8858     10.0536 -3100359680 -1273499776 -
1163001088

   meanXX  meanYY  meanZZ  StdDevXX  StdDevYY  StdDevZZ  skewXX  
skewYY  \
0  0.7355  0.7594  0.9475    0.1067    0.1166    0.1080 -0.4876 -
0.2667
1  0.6035  0.6294  0.7289    0.1152    0.1207    0.1386 -0.3131 -
0.2934
2  0.6472  0.6897  0.8902    0.0883    0.0953    0.1203 -1.9035 -
1.9296
3  0.7762  0.8041  0.9809    0.1218    0.1330    0.1301 -1.2292 -
0.9592
4  0.6365  0.6605  0.7853    0.1240    0.1293    0.1497 -0.6071 -
0.5923

   skewZZ  kurtosisXX  kurtosisYY  kurtosisZZ  entropyXX  entropyYY  \
0 -2.4022      5.4413      4.6547     13.5992  2388.4263  2222.7087   
1 -0.3768      2.9232      2.9286      2.8833  2567.9663  2531.9880   
2 -2.2574      9.4334      9.4209      9.8885  1790.9473  1715.9330   
```

```
3 -2.8063      6.0941      5.1355     12.2990  2162.7341  1909.6716
4 -0.6824      3.0376      3.0058      3.2231  2493.1421  2432.7375

   entropyZZ  ALLdaub4RR  ALLdaub4RG  ALLdaub4RB  ALLdaub4H  ALLdaub4S
\
0   512.8892    111.4315    111.9330    120.6838     0.3266     0.0390

1  2189.7100    102.9773    103.0778    106.6464     0.3270     0.0177

2   757.2745    100.8594    108.7688    117.4546     0.2938     0.0705

3   -63.9162    114.4421    114.8475    122.3142     0.3076     0.0329

4  1815.4894    105.2504    105.1742    110.4669     0.3338     0.0242


   ALLdaub4V  ALLdaub4L  ALLdaub4a  ALLdaub4b  ALLdaub4Y
ALLdaub4Cb  \
0     0.4733   113.9924    65.0610    59.5989   104.8552     67.8779

1     0.4182   105.7055    64.3685    62.2084    96.8375     65.5371

2     0.4606   109.7155    62.6423    58.7439   100.2352     68.9753

3     0.4797   116.5405    64.9069    60.2562   107.2560     67.3298

4     0.4332   107.7502    64.7071    61.3549    98.8704     66.2048


   ALLdaub4Cr  ALLdaub4XX  ALLdaub4YY  ALLdaub4ZZ    CLASS
0     63.0828      0.3673      0.3793      0.4733  Basmati
1     63.5832      0.3014      0.3144      0.3641  Arborio
2     59.8342      0.3233      0.3445      0.4448  Jasmine
3     63.2237      0.3880      0.4020      0.4904  Basmati
4     63.5378      0.3184      0.3303      0.3928  Arborio
```

```
df.shape
```

```
(75000, 107)
```

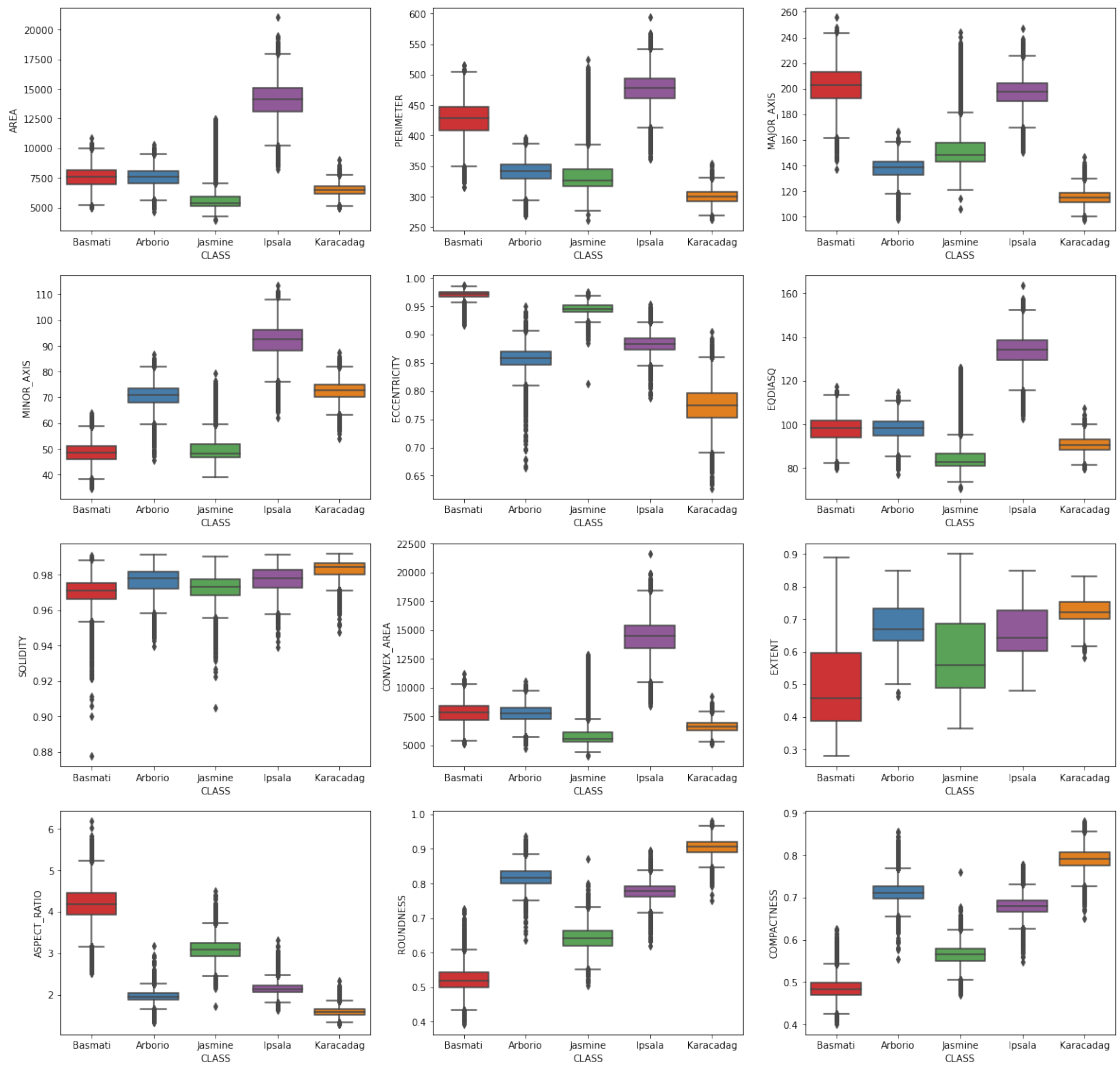
4.1. Outliers

At this point, we may also want to remove outliers. These can be due to typos in data entry, mistakes in units, or they could be legitimate but extreme values. For this project, we will remove anomalies based on the definition of extreme outliers:

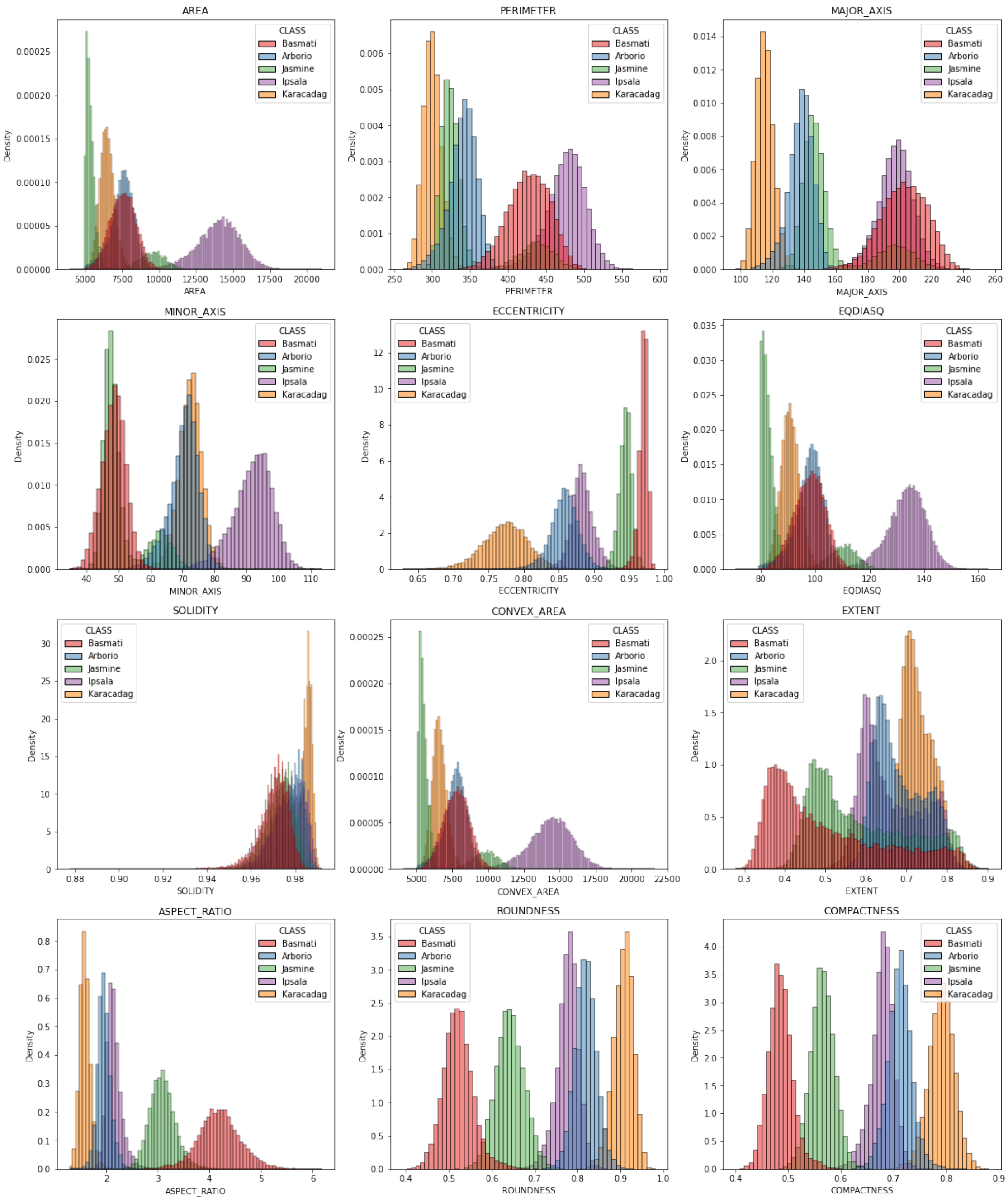
https://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm

- Below the first quartile − 3 ∗ interquartile range
- Above the third quartile + 3 ∗ interquartile range

```python
plt.figure(figsize=(20,20))
for i in range(12):
    plt.subplot(4, 3, i + 1)
    sns.boxplot(x="CLASS", y=df.columns[i], data=df, palette="Set1")
plt.show()
```




```python
plt.figure(figsize=(20,25))
for i in range(12):
    plt.subplot(4, 3, i + 1)
    sns.histplot(data=df, x=df.columns[i],hue="CLASS", stat="density",
palette="Set1")
    plt.title(df.columns[i])
plt.show()
```



4.2 Initial Data Exploration

During this stage you'll address data mining questions using querying, data visualization and reporting techniques. These may include:

- **Distribution** of key attributes (for example, the target attribute of a prediction task)
- **Relationships** between pairs or small numbers of attributes
- Results of **simple aggregations**
- **Properties** of significant sub-populations

- **Simple** statistical analyses

These analyses may directly address your data mining goals. They may also contribute to or refine the data description and quality reports, and feed into the transformation and other data preparation steps needed for further analysis.

- **Data exploration report** - Describe results of your data exploration, including first findings or initial hypothesis and their impact on the remainder of the project. If appropriate you could include graphs and plots here to indicate data characteristics that suggest further examination of interesting data subsets.

4.2.1 Distributions

```
df.describe()

               AREA     PERIMETER    MAJOR_AXIS    MINOR_AXIS
ECCENTRICITY  \
count  75000.000000  75000.000000  75000.000000  75000.000000
75000.000000
mean    8379.197507    378.169453    161.805540     66.829335
0.886077
std     3119.209274     70.597008     36.461005     16.689269
0.071906
min     3929.000000    261.040000     96.968300     34.673000
0.627700
25%     6259.000000    316.431500    132.623500     49.650200
0.846100
50%     7345.000000    351.261000    149.343950     69.183900
0.885600
75%     8901.000000    444.986000    197.462025     75.814125
0.950800
max    21019.000000    593.698000    255.647200    113.441100
0.986800

            EQDIASQ      SOLIDITY   CONVEX_AREA        EXTENT
ASPECT_RATIO  \
count  75000.000000  75000.000000  75000.000000  75000.000000
75000.000000
mean     101.731251      0.975896   8584.862320      0.633226
2.597063
std       17.874070      0.007966   3189.298025      0.123795
0.968982
min       70.728800      0.877500   4032.000000      0.278800
1.284500
25%       89.270400      0.970900   6385.000000      0.561000
1.876100
50%       96.705500      0.976400   7532.000000      0.655800
2.153200
75%      106.457100      0.982200   9153.000000      0.727800
3.228700
max      163.591600      0.992100  21633.000000      0.901700
```

6.179500

	ROUNDNESS	COMPACTNESS	SHAPEFACTOR_1	SHAPEFACTOR_2
count	75000.000000	75000.000000	75000.000000	75000.000000
mean	0.732505	0.646079	0.020619	0.008407
std	0.138637	0.110787	0.005287	0.001903
min	0.392500	0.400600	0.011300	0.005100
25%	0.620600	0.551100	0.017000	0.006600
50%	0.775400	0.677100	0.018600	0.008700
75%	0.834500	0.725300	0.026200	0.009700
max	0.980000	0.879900	0.036900	0.013500

	SHAPEFACTOR_3	SHAPEFACTOR_4	meanRR	meanRG	meanRB
count	75000.000000	75000.000000	75000.000000	75000.000000	75000.000000
mean	0.429692	0.985509	216.398005	218.205782	227.918353
std	0.141146	0.007280	13.308330	13.646445	10.682523
min	0.160500	0.896200	153.800000	157.249900	160.158400
25%	0.303700	0.981600	206.605125	207.848625	220.927750
50%	0.458500	0.986400	215.118800	217.137550	228.801250
75%	0.526100	0.990700	225.016125	227.339300	236.171600
max	0.774300	0.999000	252.183700	252.323100	252.108500

	StdDevRR	StdDevRG	StdDevRB	skewRR	skewRG
count	75000.000000	75000.000000	75000.000000	75000.000000	75000.000000
mean	15.342766	15.449838	15.477779	-1.778549	-1.938456
std	3.454178	3.562578	3.468618	0.948735	1.111904
min	6.817100	6.411700	6.417500	-6.938800	-7.911800
25%	12.579400	12.741500	13.050675	-2.360500	-2.522300
50%	15.542900	15.686150	15.539300	-1.608600	-1.683350
75%	17.881000	18.032225	17.891800	-1.095000	-1.136700
max	29.967400	30.765400	30.858000	0.917900	0.771900

```
             skewRB    kurtosisRR    kurtosisRG    kurtosisRB
entropyRR  \
count  75000.000000  75000.000000  75000.000000  75000.000000
7.500000e+04
mean      -2.360081     11.955533     12.944259     14.467290 -
4.426130e+09
std        0.950987      7.479528      9.302984      7.754649
2.239719e+09
min       -6.938200      1.841300      1.878100      1.885200 -
1.356042e+10
25%       -3.002300      6.481900      6.536450      8.425450 -
4.604666e+09
50%       -2.321100      9.727700     10.003650     13.436350 -
3.626555e+09
75%       -1.609700     15.068550     15.746325     18.253800 -
2.877646e+09
max        1.162400     75.201600     89.363100     71.980400 -
1.474496e+09

          entropyRG     entropyRB         meanH         meanS
meanV  \
count  7.500000e+04  7.500000e+04  75000.000000  75000.000000
75000.000000
mean  -4.509678e+09 -4.820370e+09      0.547699      0.060556
0.898100
std    2.268614e+09  1.994516e+09      0.185565      0.036708
0.043447
min   -1.383477e+10 -1.317801e+10      0.034100      0.001400
0.628100
25%   -4.870068e+09 -5.477346e+09      0.526275      0.027200
0.868600
50%   -3.674473e+09 -4.153602e+09      0.644800      0.054250
0.903500
75%   -2.890273e+09 -3.415980e+09      0.664600      0.091700
0.932300
max   -1.550952e+09 -1.605495e+09      0.817100      0.241700
0.990600

            StdDevH      StdDevS       StdDevV         skewH
skewS  \
count  75000.000000  75000.000000  75000.000000  75000.000000
75000.000000
mean       0.064218      0.019138      0.060251     -4.797680
0.019438
std        0.061397      0.010459      0.013644      7.194686
1.043360
min        0.002100      0.003000      0.025100    -70.866500     -
2.713400
25%        0.022200      0.010900      0.050500     -8.312225     -
0.743300
```

```
50%        0.041700      0.015800      0.060700     -3.579100     -
0.277450
75%        0.087425      0.025200      0.069900     -0.169475
0.810825
max        0.410300      0.093900      0.118400     25.021800
6.927700

              skewV     kurtosisH     kurtosisS     kurtosisV
entropyH  \
count  75000.000000  75000.000000  75000.000000  75000.000000
75000.000000
mean      -2.489326    131.841205      4.601397     15.402438
2309.978182
std        1.051619    261.985126      2.819120      9.049990
602.876603
min       -7.910400      1.009300      1.275100      1.885000
262.201600
25%       -3.149800      8.462025      2.874975      8.655600
1944.970550
50%       -2.428750     41.961300      4.007750     13.850600
2338.881600
75%       -1.659675    146.237675      5.489300     19.087700
2713.329300
max        0.777400   5504.557100     76.759400     89.212900
4868.357900

           entropyS      entropyV         meanL         meanA
meanB  \
count  75000.000000  75000.000000  75000.000000  75000.000000
75000.000000
mean     184.779482   1246.241590    222.215488    128.759108
122.920756
std      159.467177    468.043347     11.801014      2.352168
4.873599
min        1.619600    137.279200    164.704200    118.268500
107.303700
25%       52.218900    910.651125    213.312575    127.767675
118.926575
50%      123.785950   1194.487450    221.581450    128.838500
122.847600
75%      291.221550   1518.092750    230.164650    130.348625
126.154775
max      975.833900   4814.083500    252.505700    134.923800
140.567600

            StdDevL       StdDevA       StdDevB         skewL
skewA  \
count  75000.000000  75000.000000  75000.000000  75000.000000
75000.000000
mean      14.127218      0.939870      2.215095     -2.083832
```

```
0.114552
std       3.247309      0.395952      1.327015      1.072309
0.914165
min       5.840700      0.106100      0.000000     -7.911300     -
8.295200
25%      11.702575      0.629200      1.177900     -2.675600     -
0.467425
50%      14.405750      0.818300      1.674700     -1.863400
0.039600
75%      16.445050      1.220200      3.043725     -1.296075
0.756100
max      27.440700      3.258200     10.821100      0.671300
9.208500

              skewB     kurtosisL     kurtosisA     kurtosisB
entropyL  \
count  75000.000000  75000.000000  75000.000000  75000.000000
7.500000e+04
mean       0.529561     13.850684      4.282437      4.731105 -
4.654252e+09
std        0.997268      9.346636      2.361591      2.969105
2.248731e+09
min       -3.168200      1.918700      1.000000      0.999900 -
1.389150e+10
25%       -0.049300      7.333000      2.800700      2.905700 -
4.995975e+09
50%        0.581650     11.042450      3.702300      3.927850 -
3.842979e+09
75%        1.081900     16.932775      5.058150      5.611525 -
3.060054e+09
max        8.540500     89.816000     85.788600     73.882000 -
1.690136e+09

           entropyA      entropyB         meanY        meanCb
meanCr  \
count  7.500000e+04  7.500000e+04  75000.000000  75000.000000
75000.000000
mean  -1.342112e+09 -1.248950e+09    203.886710    132.483100
126.403359
std    4.750043e+08  5.704776e+08     10.866061      4.320342
2.350780
min   -3.267201e+09 -3.472658e+09    150.474500    116.642000
114.724000
25%   -1.404111e+09 -1.305433e+09    195.709775    129.684175
126.104525
50%   -1.196488e+09 -1.037515e+09    203.406400    132.584850
127.043700
75%   -1.035494e+09 -8.441662e+08    211.211900    136.026325
127.467425
max   -6.293092e+08 -5.472335e+08    232.550000    146.855400
```

```
133.076200

             StdDevY      StdDevCb      StdDevCr         skewY
skewCb  \
count  75000.000000  75000.000000  75000.000000  75000.000000
75000.000000
mean      13.163794      1.969016      0.754401     -1.955500     -
0.471512
std        2.979435      1.154209      0.407306      1.014331
1.227026
min        5.589200      0.000000      0.000000     -7.531800     -
7.501200
25%       10.892900      1.078000      0.468500     -2.521600     -
1.080800
50%       13.379150      1.511050      0.687700     -1.750650     -
0.586200
75%       15.288100      2.700025      0.923100     -1.210200
0.100025
max       25.460100      9.474100      4.167300      0.866100
120.657500

              skewCr     kurtosisY    kurtosisCb    kurtosisCr
entropyY  \
count  75000.000000  75000.000000  75000.000000  75000.000000
7.500000e+04
mean       1.734350     12.894485      5.129198     67.688294 -
3.856125e+09
std        7.868671      8.536443     58.357386    560.170304
1.862941e+09
min       -9.581300      1.871700      1.000000      0.999900 -
1.152219e+10
25%        0.140775      6.914600      2.847700      2.638800 -
4.134462e+09
50%        0.606200     10.370950      3.959400      3.572000 -
3.187489e+09
75%        1.343700     15.792500      5.734350      5.356425 -
2.539634e+09
max      116.318200     83.392400  14559.161100  13529.946300 -
1.380124e+09

           entropyCb     entropyCr        meanXX        meanYY
meanZZ  \
count  7.500000e+04  7.500000e+04  75000.000000  75000.000000
75000.000000
mean  -1.414347e+09 -1.300141e+09      0.684125      0.714363
0.842659
std    4.388745e+08  5.041548e+08      0.083806      0.094242
0.086255
min   -3.219491e+09 -3.294497e+09      0.319800      0.338400
0.384200
```

```
25%   -1.578173e+09 -1.336365e+09      0.620200      0.642100
0.782700
50%   -1.288827e+09 -1.134926e+09      0.680400      0.706700
0.848700
75%   -1.126542e+09 -9.659853e+08      0.740700      0.776000
0.909600
max   -6.653100e+08 -5.999740e+08      0.927600      0.977000
1.061300

           StdDevXX      StdDevYY      StdDevZZ       skewXX
skewYY  \
count  75000.000000  75000.000000  75000.000000  75000.00000
75000.000000
mean       0.097866      0.103307      0.116325     -1.15768     -
1.131194
std        0.021346      0.023054      0.025187      0.82280
0.900195
min        0.048300      0.049700      0.053300     -5.72170     -
6.170300
25%        0.079700      0.083600      0.096700     -1.60470     -
1.590575
50%        0.097200      0.102000      0.116600     -0.99310     -
0.932550
75%        0.113400      0.120500      0.133100     -0.58500     -
0.510000
max        0.194900      0.211900      0.226200      1.68760
1.633800

             skewZZ    kurtosisXX    kurtosisYY    kurtosisZZ
entropyXX  \
count  75000.000000  75000.000000  75000.000000  75000.000000
75000.000000
mean      -1.504015      8.279693      8.306708      9.274027
2588.904041
std        0.829523      5.298437      5.769769      5.151446
743.477905
min       -5.953800      1.693700      1.698000      1.691000
764.255100
25%       -2.028425      4.476275      4.255000      5.220350
2089.496050
50%       -1.416100      6.707100      6.432400      8.568450
2395.386350
75%       -0.883975     10.304150     10.273325     11.667725
2805.124300
max        1.864400     53.962900     58.776300     49.425100
6322.974600

          entropyYY     entropyZZ    ALLdaub4RR    ALLdaub4RG
ALLdaub4RB  \
count  75000.000000  75000.000000  75000.000000  75000.000000
```

```
75000.000000
mean    2367.177736   1489.687843    108.178754    109.082089
113.936257
std      596.172238    828.002085      6.657980      6.827630
5.343310
min      343.706900  -2074.590800     76.843600     78.572300
80.027800
25%     1975.926700    963.928125    103.278675    103.900075
110.438275
50%     2268.014150   1481.069350    107.534300    108.545600
114.379400
75%     2574.653375   1924.440875    112.486075    113.654875
118.063450
max     5835.086900   5615.509800    126.105600    126.169700
126.067200

          ALLdaub4H     ALLdaub4S     ALLdaub4V     ALLdaub4L
ALLdaub4a  \
count  75000.000000  75000.000000  75000.000000  75000.000000
75000.000000
mean       0.273845      0.030271      0.448960    111.088252
64.379443
std        0.092785      0.018347      0.021736      5.904854
1.175616
min        0.017200      0.000700      0.313900     82.300600
59.137900
25%        0.263100      0.013600      0.434200    106.632900
63.883800
50%        0.322400      0.027100      0.451600    110.770700
64.419350
75%        0.332300      0.045800      0.466100    115.065075
65.174200
max        0.408700      0.120800      0.495100    126.265100
67.459000

          ALLdaub4b     ALLdaub4Y    ALLdaub4Cb    ALLdaub4Cr
ALLdaub4XX  \
count  75000.000000  75000.000000  75000.000000  75000.000000
75000.000000
mean      61.461457    101.925425     66.240541     63.202088
0.341944
std        2.435635      5.436861      2.159109      1.174976
0.041921
min       53.653800     75.191800     58.323800     57.363400
0.159700
25%       59.465575     97.834400     64.842000     63.052800
0.309900
50%       61.424400    101.683700     66.291600     63.522050
0.340100
75%       63.076825    105.592450     68.011800     63.734000
```

```
0.370300
max       70.284000    116.287300     73.424700     66.539100
0.463900

         ALLdaub4YY    ALLdaub4ZZ
count  75000.000000  75000.000000
mean       0.357058      0.421176
std        0.047139      0.043137
min        0.169000      0.191800
25%        0.320900      0.391200
50%        0.353300      0.424200
75%        0.387900      0.454700
max        0.488600      0.530200
```

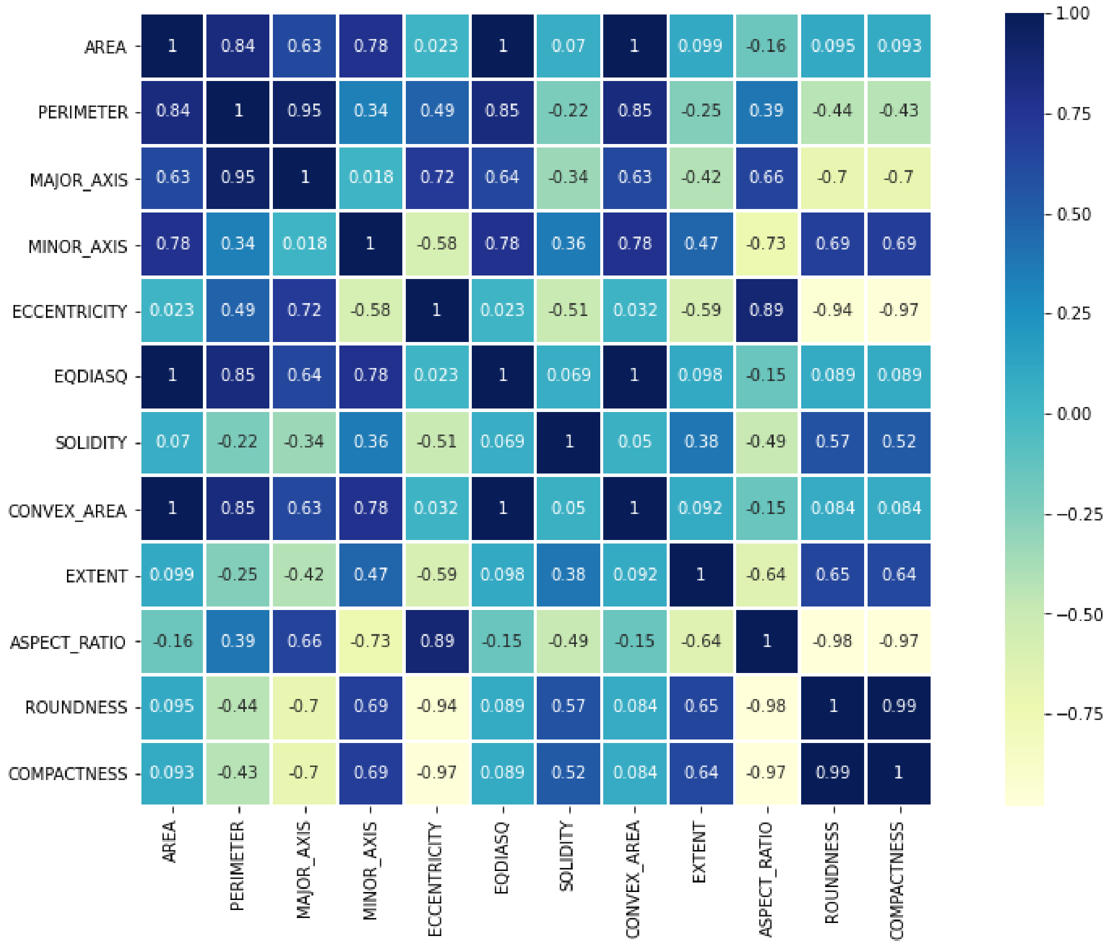
4.2.2 Correlations

Can we derive any correlation from this data-set. Pairplot chart gives us correlations, distributions and regression path Correlogram are awesome for exploratory analysis. It allows to quickly observe the relationship between every variable of your matrix. It is easy to do it with seaborn: just call the pairplot function

Pairplot Documentation cab be found here: https://seaborn.pydata.org/generated/seaborn.pairplot.html

```
plt.figure(figsize=(16,9))
sns.heatmap(df.iloc[:,:12].corr(), cmap="YlGnBu",annot=True,
fmt=".2g", linewidths = 1, square= True)
```

```
<AxesSubplot:>
```



4.3 Data Quality Report

List the results of the data quality verification. If quality problems exist, suggest possible solutions. Solutions to data quality problems generally depend heavily on both data and business knowledge.

- Primary data quality issue is missing values in certain parts of the data. To correct for this, we imputed the mean value of the columns into those missing fields in order to have a more complete approximation of the missing data.

5. Stage Four - Modelling

As the first step in modelling, you'll select the actual modelling technique that you'll be using. Although you may have already selected a tool during the business understanding phase, at this stage you'll be selecting the specific modelling technique e.g. decision-tree building with C5.0, or neural network generation with back propagation. If multiple techniques are applied, perform this task separately for each technique.

5.1. Modelling technique

Document the actual modelling technique that is to be used.

Import Models below:

5.2. Modelling assumptions

Many modelling techniques make specific assumptions about the data, for example that all attributes have uniform distributions, no missing values allowed, class attribute must be symbolic etc. Record any assumptions made.

5.3. Build Model

Run the modelling tool on the prepared dataset to create one or more models.

Parameter settings - With any modelling tool there are often a large number of parameters that can be adjusted. List the parameters and their chosen values, along with the rationale for the choice of parameter settings.

Models - These are the actual models produced by the modelling tool, not a report on the models.

Model descriptions - Describe the resulting models, report on the interpretation of the models and document any difficulties encountered with their meanings.

5.4. Assess Model

Interpret the models according to your domain knowledge, your data mining success criteria and your desired test design. Judge the success of the application of modelling and discovery techniques technically, then contact business analysts and domain experts later in order to discuss the data mining results in the business context. This task only considers models, whereas the evaluation phase also takes into account all other results that were produced in the course of the project.

At this stage you should rank the models and assess them according to the evaluation criteria. You should take the business objectives and business success criteria into account as far as you can here. In most data mining projects a single technique is applied more than once and data mining results are generated with several different techniques.

Model assessment - Summarise the results of this task, list the qualities of your generated models (e.g.in terms of accuracy) and rank their quality in relation to each other.

Revised parameter settings - According to the model assessment, revise parameter settings and tune them for the next modelling run. Iterate model building and assessment until you strongly believe that you have found the best model(s). Document all such revisions and assessments.