

Beans

December 16, 2022

```
In [65]: import pandas as pd
from sklearn.model_selection import train_test_split
import seaborn as sns
from sklearn.ensemble import RandomForestClassifier
import matplotlib.pyplot as plt

df = pd.read_csv('/public/bmort/python/beans.csv')
print(df.isnull().sum())
print("There is missing data in the ShapeFactor3 column so lets fix this")

imputed_value = df['ShapeFactor3'].median()
df['ShapeFactor3'].fillna(imputed_value)
df['ShapeFactor3'] = df['ShapeFactor3'].fillna(imputed_value)

print("")

print(df.describe())

print("The magnitudes between the columns are much different, some are very small and some are very large")
print("The range between the columns varies a fair amount as well with some exceptions")
print("The range of area is large, the range of the convex area is also very large")
print("It is clear to see there are outliers here because there is the max is so far off")

sns.heatmap(df.corr(), cmap="YlGnBu")
plt.show()
print(df.corr())

print("")

train_x = df[['Area', 'Perimeter', 'AspectRatio', 'Eccentricity', 'roundness', 'Compactness', 'Convexity', 'ConvexArea', 'ConvexPerim', 'ConvexRatio', 'Convexity2', 'Convexity3', 'Convexity4', 'Convexity5', 'Convexity6', 'Convexity7', 'Convexity8', 'Convexity9', 'Convexity10', 'Convexity11', 'Convexity12', 'Convexity13', 'Convexity14', 'Convexity15', 'Convexity16', 'Convexity17', 'Convexity18', 'Convexity19', 'Convexity20', 'Convexity21', 'Convexity22', 'Convexity23', 'Convexity24', 'Convexity25', 'Convexity26', 'Convexity27', 'Convexity28', 'Convexity29', 'Convexity30', 'Convexity31', 'Convexity32', 'Convexity33', 'Convexity34', 'Convexity35', 'Convexity36', 'Convexity37', 'Convexity38', 'Convexity39', 'Convexity40', 'Convexity41', 'Convexity42', 'Convexity43', 'Convexity44', 'Convexity45', 'Convexity46', 'Convexity47', 'Convexity48', 'Convexity49', 'Convexity50', 'Convexity51', 'Convexity52', 'Convexity53', 'Convexity54', 'Convexity55', 'Convexity56', 'Convexity57', 'Convexity58', 'Convexity59', 'Convexity60', 'Convexity61', 'Convexity62', 'Convexity63', 'Convexity64', 'Convexity65', 'Convexity66', 'Convexity67', 'Convexity68', 'Convexity69', 'Convexity70', 'Convexity71', 'Convexity72', 'Convexity73', 'Convexity74', 'Convexity75', 'Convexity76', 'Convexity77', 'Convexity78', 'Convexity79', 'Convexity80', 'Convexity81', 'Convexity82', 'Convexity83', 'Convexity84', 'Convexity85', 'Convexity86', 'Convexity87', 'Convexity88', 'Convexity89', 'Convexity90', 'Convexity91', 'Convexity92', 'Convexity93', 'Convexity94', 'Convexity95', 'Convexity96', 'Convexity97', 'Convexity98', 'Convexity99', 'Convexity100', 'Convexity101', 'Convexity102', 'Convexity103', 'Convexity104', 'Convexity105', 'Convexity106', 'Convexity107', 'Convexity108', 'Convexity109', 'Convexity110', 'Convexity111', 'Convexity112', 'Convexity113', 'Convexity114', 'Convexity115', 'Convexity116', 'Convexity117', 'Convexity118', 'Convexity119', 'Convexity120', 'Convexity121', 'Convexity122', 'Convexity123', 'Convexity124', 'Convexity125', 'Convexity126', 'Convexity127', 'Convexity128', 'Convexity129', 'Convexity130', 'Convexity131', 'Convexity132', 'Convexity133', 'Convexity134', 'Convexity135', 'Convexity136', 'Convexity137', 'Convexity138', 'Convexity139', 'Convexity140', 'Convexity141', 'Convexity142', 'Convexity143', 'Convexity144', 'Convexity145', 'Convexity146', 'Convexity147', 'Convexity148', 'Convexity149', 'Convexity150', 'Convexity151', 'Convexity152', 'Convexity153', 'Convexity154', 'Convexity155', 'Convexity156', 'Convexity157', 'Convexity158', 'Convexity159', 'Convexity160', 'Convexity161', 'Convexity162', 'Convexity163', 'Convexity164', 'Convexity165', 'Convexity166', 'Convexity167', 'Convexity168', 'Convexity169', 'Convexity170', 'Convexity171', 'Convexity172', 'Convexity173', 'Convexity174', 'Convexity175', 'Convexity176', 'Convexity177', 'Convexity178', 'Convexity179', 'Convexity180', 'Convexity181', 'Convexity182', 'Convexity183', 'Convexity184', 'Convexity185', 'Convexity186', 'Convexity187', 'Convexity188', 'Convexity189', 'Convexity190', 'Convexity191', 'Convexity192', 'Convexity193', 'Convexity194', 'Convexity195', 'Convexity196', 'Convexity197', 'Convexity198', 'Convexity199', 'Convexity200', 'Convexity201', 'Convexity202', 'Convexity203', 'Convexity204', 'Convexity205', 'Convexity206', 'Convexity207', 'Convexity208', 'Convexity209', 'Convexity210', 'Convexity211', 'Convexity212', 'Convexity213', 'Convexity214', 'Convexity215', 'Convexity216', 'Convexity217', 'Convexity218', 'Convexity219', 'Convexity220', 'Convexity221', 'Convexity222', 'Convexity223', 'Convexity224', 'Convexity225', 'Convexity226', 'Convexity227', 'Convexity228', 'Convexity229', 'Convexity230', 'Convexity231', 'Convexity232', 'Convexity233', 'Convexity234', 'Convexity235', 'Convexity236', 'Convexity237', 'Convexity238', 'Convexity239', 'Convexity240', 'Convexity241', 'Convexity242', 'Convexity243', 'Convexity244', 'Convexity245', 'Convexity246', 'Convexity247', 'Convexity248', 'Convexity249', 'Convexity250', 'Convexity251', 'Convexity252', 'Convexity253', 'Convexity254', 'Convexity255', 'Convexity256', 'Convexity257', 'Convexity258', 'Convexity259', 'Convexity260', 'Convexity261', 'Convexity262', 'Convexity263', 'Convexity264', 'Convexity265', 'Convexity266', 'Convexity267', 'Convexity268', 'Convexity269', 'Convexity270', 'Convexity271', 'Convexity272', 'Convexity273', 'Convexity274', 'Convexity275', 'Convexity276', 'Convexity277', 'Convexity278', 'Convexity279', 'Convexity280', 'Convexity281', 'Convexity282', 'Convexity283', 'Convexity284', 'Convexity285', 'Convexity286', 'Convexity287', 'Convexity288', 'Convexity289', 'Convexity290', 'Convexity291', 'Convexity292', 'Convexity293', 'Convexity294', 'Convexity295', 'Convexity296', 'Convexity297', 'Convexity298', 'Convexity299', 'Convexity300', 'Convexity301', 'Convexity302', 'Convexity303', 'Convexity304', 'Convexity305', 'Convexity306', 'Convexity307', 'Convexity308', 'Convexity309', 'Convexity310', 'Convexity311', 'Convexity312', 'Convexity313', 'Convexity314', 'Convexity315', 'Convexity316', 'Convexity317', 'Convexity318', 'Convexity319', 'Convexity320', 'Convexity321', 'Convexity322', 'Convexity323', 'Convexity324', 'Convexity325', 'Convexity326', 'Convexity327', 'Convexity328', 'Convexity329', 'Convexity330', 'Convexity331', 'Convexity332', 'Convexity333', 'Convexity334', 'Convexity335', 'Convexity336', 'Convexity337', 'Convexity338', 'Convexity339', 'Convexity340', 'Convexity341', 'Convexity342', 'Convexity343', 'Convexity344', 'Convexity345', 'Convexity346', 'Convexity347', 'Convexity348', 'Convexity349', 'Convexity350', 'Convexity351', 'Convexity352', 'Convexity353', 'Convexity354', 'Convexity355', 'Convexity356', 'Convexity357', 'Convexity358', 'Convexity359', 'Convexity360', 'Convexity361', 'Convexity362', 'Convexity363', 'Convexity364', 'Convexity365', 'Convexity366', 'Convexity367', 'Convexity368', 'Convexity369', 'Convexity370', 'Convexity371', 'Convexity372', 'Convexity373', 'Convexity374', 'Convexity375', 'Convexity376', 'Convexity377', 'Convexity378', 'Convexity379', 'Convexity380', 'Convexity381', 'Convexity382', 'Convexity383', 'Convexity384', 'Convexity385', 'Convexity386', 'Convexity387', 'Convexity388', 'Convexity389', 'Convexity390', 'Convexity391', 'Convexity392', 'Convexity393', 'Convexity394', 'Convexity395', 'Convexity396', 'Convexity397', 'Convexity398', 'Convexity399', 'Convexity400', 'Convexity401', 'Convexity402', 'Convexity403', 'Convexity404', 'Convexity405', 'Convexity406', 'Convexity407', 'Convexity408', 'Convexity409', 'Convexity410', 'Convexity411', 'Convexity412', 'Convexity413', 'Convexity414', 'Convexity415', 'Convexity416', 'Convexity417', 'Convexity418', 'Convexity419', 'Convexity420', 'Convexity421', 'Convexity422', 'Convexity423', 'Convexity424', 'Convexity425', 'Convexity426', 'Convexity427', 'Convexity428', 'Convexity429', 'Convexity430', 'Convexity431', 'Convexity432', 'Convexity433', 'Convexity434', 'Convexity435', 'Convexity436', 'Convexity437', 'Convexity438', 'Convexity439', 'Convexity440', 'Convexity441', 'Convexity442', 'Convexity443', 'Convexity444', 'Convexity445', 'Convexity446', 'Convexity447', 'Convexity448', 'Convexity449', 'Convexity450', 'Convexity451', 'Convexity452', 'Convexity453', 'Convexity454', 'Convexity455', 'Convexity456', 'Convexity457', 'Convexity458', 'Convexity459', 'Convexity460', 'Convexity461', 'Convexity462', 'Convexity463', 'Convexity464', 'Convexity465', 'Convexity466', 'Convexity467', 'Convexity468', 'Convexity469', 'Convexity470', 'Convexity471', 'Convexity472', 'Convexity473', 'Convexity474', 'Convexity475', 'Convexity476', 'Convexity477', 'Convexity478', 'Convexity479', 'Convexity480', 'Convexity481', 'Convexity482', 'Convexity483', 'Convexity484', 'Convexity485', 'Convexity486', 'Convexity487', 'Convexity488', 'Convexity489', 'Convexity
```

```

df['le_class'] = le.transform(df['Class'])

train_x = df[['MajorAxisLength', 'MinorAxisLength', 'AspectRatio', 'Extent', 'Solidity', 'Roundness', 'Compactness', 'ShapeFactor1', 'ShapeFactor2', 'ShapeFactor3', 'ShapeFactor4']]
train_y = df['le_class'].values
X_train, X_test, y_train, y_test = train_test_split(train_x, train_y, test_size=0.2)

rf = RandomForestClassifier(n_estimators = 50)
rf.fit(X_train, y_train);

y_pred = rf.predict(X_test)
print(rf.score(X_train, y_train))
print(rf.score(X_test, y_test))

from sklearn.metrics import confusion_matrix
confusion_matrix(y_test, y_pred)

from sklearn.metrics import ConfusionMatrixDisplay
cm = confusion_matrix(y_test, y_pred)
disp = ConfusionMatrixDisplay(confusion_matrix=cm)
disp.plot()

bean = pd.read_csv('/public/bmort/python/beans-unknown.csv')
print(bean)
test_bean = bean[['MajorAxisLength', 'MinorAxisLength', 'AspectRatio', 'Extent', 'Solidity', 'Roundness', 'Compactness', 'ShapeFactor1', 'ShapeFactor2', 'ShapeFactor3', 'ShapeFactor4']]
bean_pred = rf.predict(test_bean)
print("This is the bean prediction of the sample data")
print(bean_pred)

```

```

Area          0
Perimeter     0
MajorAxisLength  0
MinorAxisLength  0
AspectRatio    0
Eccentricity   0
ConvexArea     0
EquivDiameter  0
Extent         0
Solidity       0
roundness      0
Compactness    0
ShapeFactor1   0
ShapeFactor2   0
ShapeFactor3    1
ShapeFactor4   0
Class          0
dtype: int64

```

There is missing data in the ShapeFactor3 column so lets fix this

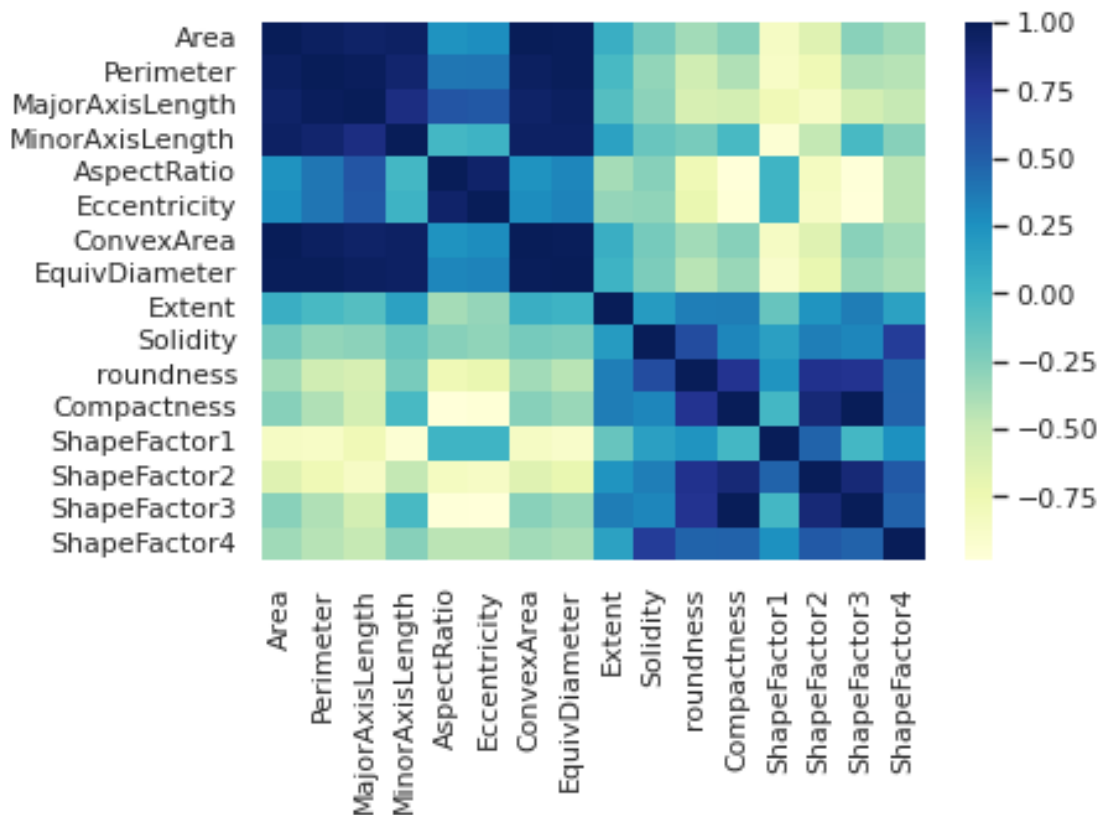
	Area	Perimeter	MajorAxisLength	MinorAxisLength	\
count	13533.000000	13533.000000	13533.000000	13533.000000	
mean	53057.388384	855.066479	319.922981	202.378468	
std	29401.235132	214.781993	85.833897	45.064896	
min	20420.000000	524.736000	183.601165	122.512653	
25%	36269.000000	703.180000	253.059398	175.884179	
50%	44581.000000	793.897000	296.441382	192.500481	
75%	61422.000000	977.266000	376.352986	217.263451	
max	254616.000000	1985.370000	738.860154	460.198497	

	AspectRatio	Eccentricity	ConvexArea	EquivDiameter	Extent	\
count	13533.000000	13533.000000	13533.000000	13533.000000	13533.000000	
mean	1.581111	0.750310	53777.120890	253.052530	0.749827	
std	0.245328	0.091890	29853.152141	59.324886	0.048939	
min	1.024868	0.218951	20684.000000	161.243764	0.555315	
25%	1.430641	0.715134	36669.000000	214.893288	0.718753	
50%	1.549898	0.764011	45123.000000	238.248383	0.759903	
75%	1.704026	0.809699	62388.000000	279.651425	0.786847	
max	2.430306	0.911423	263261.000000	569.374358	0.866195	

	Solidity	roundness	Compactness	ShapeFactor1	ShapeFactor2	\
count	13533.000000	13533.000000	13533.000000	13533.000000	13533.000000	
mean	0.987150	0.873653	0.800347	0.006561	0.001719	
std	0.004651	0.059410	0.061485	0.001130	0.000596	
min	0.919246	0.489618	0.640577	0.002778	0.000564	
25%	0.985675	0.833360	0.763181	0.005890	0.001158	
50%	0.988287	0.883447	0.801505	0.006643	0.001700	
75%	0.990018	0.917039	0.834520	0.007271	0.002173	
max	0.994677	0.990685	0.987303	0.010451	0.003665	

	ShapeFactor3	ShapeFactor4
count	13533.000000	13533.000000
mean	0.644336	0.995077
std	0.098687	0.004348
min	0.410339	0.947687
25%	0.582445	0.993717
50%	0.642410	0.996393
75%	0.696423	0.997891
max	0.974767	0.999733

The magnitudes between the columns are much different, some are very small and some are large
The range between the columns varies a fair amount as well with some exceptions where they are
The range of area is large, the range of the convex area is also very large
It is clear to see there are outliers here because there is the max is so far off of the mean.



	Area	Perimeter	MajorAxisLength	MinorAxisLength	\
Area	1.000000	0.966904	0.932615	0.952038	
Perimeter	0.966904	1.000000	0.977558	0.914326	
MajorAxisLength	0.932615	0.977558	1.000000	0.828341	
MinorAxisLength	0.952038	0.914326	0.828341	1.000000	
AspectRatio	0.243660	0.386073	0.550062	-0.005404	
Eccentricity	0.268623	0.391125	0.541075	0.022423	
ConvexArea	0.999940	0.967868	0.933384	0.951777	
EquivDiameter	0.984997	0.991452	0.962267	0.949208	
Extent	0.054675	-0.020630	-0.077350	0.146002	
Solidity	-0.197216	-0.304551	-0.284758	-0.156688	
roundness	-0.358979	-0.548265	-0.595651	-0.213982	
Compactness	-0.269787	-0.407432	-0.567913	-0.018598	
ShapeFactor1	-0.848382	-0.865748	-0.775824	-0.947191	
ShapeFactor2	-0.641205	-0.768603	-0.859415	-0.475313	
ShapeFactor3	-0.273756	-0.408907	-0.567630	-0.022736	
ShapeFactor4	-0.357928	-0.431119	-0.484385	-0.266295	

	AspectRatio	Eccentricity	ConvexArea	EquivDiameter	\
Area	0.243660	0.268623	0.999940	0.984997	
Perimeter	0.386073	0.391125	0.967868	0.991452	

MajorAxisLength	0.550062	0.541075	0.933384	0.962267
MinorAxisLength	-0.005404	0.022423	0.951777	0.949208
AspectRatio	1.000000	0.924207	0.245229	0.305206
Eccentricity	0.924207	1.000000	0.270393	0.319410
ConvexArea	0.245229	0.270393	1.000000	0.985254
EquivDiameter	0.305206	0.319410	0.985254	1.000000
Extent	-0.371479	-0.319910	0.052892	0.028773
Solidity	-0.269104	-0.298372	-0.206784	-0.232230
roundness	-0.764988	-0.720220	-0.363531	-0.437107
Compactness	-0.987647	-0.970317	-0.271641	-0.328977
ShapeFactor1	0.020914	0.017238	-0.848374	-0.893397
ShapeFactor2	-0.837337	-0.859269	-0.642770	-0.714696
ShapeFactor3	-0.978534	-0.981064	-0.275634	-0.331603
ShapeFactor4	-0.451580	-0.450671	-0.364211	-0.394600

	Extent	Solidity	roundness	Compactness	ShapeFactor1 \
Area	0.054675	-0.197216	-0.358979	-0.269787	-0.848382
Perimeter	-0.020630	-0.304551	-0.548265	-0.407432	-0.865748
MajorAxisLength	-0.077350	-0.284758	-0.595651	-0.567913	-0.775824
MinorAxisLength	0.146002	-0.156688	-0.213982	-0.018598	-0.947191
AspectRatio	-0.371479	-0.269104	-0.764988	-0.987647	0.020914
Eccentricity	-0.319910	-0.298372	-0.720220	-0.970317	0.017238
ConvexArea	0.052892	-0.206784	-0.363531	-0.271641	-0.848374
EquivDiameter	0.028773	-0.232230	-0.437107	-0.328977	-0.893397
Extent	1.000000	0.192236	0.344658	0.355158	-0.141615
Solidity	0.192236	1.000000	0.609621	0.304833	0.154229
roundness	0.344658	0.609621	1.000000	0.766030	0.234064
Compactness	0.355158	0.304833	0.766030	1.000000	-0.005994
ShapeFactor1	-0.141615	0.154229	0.234064	-0.005994	1.000000
ShapeFactor2	0.237762	0.344337	0.781478	0.868350	0.473225
ShapeFactor3	0.348469	0.308660	0.761057	0.998685	-0.005055
ShapeFactor4	0.148651	0.700132	0.472688	0.486344	0.251063

	ShapeFactor2	ShapeFactor3	ShapeFactor4
Area	-0.641205	-0.273756	-0.357928
Perimeter	-0.768603	-0.408907	-0.431119
MajorAxisLength	-0.859415	-0.567630	-0.484385
MinorAxisLength	-0.475313	-0.022736	-0.266295
AspectRatio	-0.837337	-0.978534	-0.451580
Eccentricity	-0.859269	-0.981064	-0.450671
ConvexArea	-0.642770	-0.275634	-0.364211
EquivDiameter	-0.714696	-0.331603	-0.394600
Extent	0.237762	0.348469	0.148651
Solidity	0.344337	0.308660	0.700132
roundness	0.781478	0.761057	0.472688
Compactness	0.868350	0.998685	0.486344
ShapeFactor1	0.473225	-0.005055	0.251063
ShapeFactor2	1.000000	0.872323	0.531714

ShapeFactor3	0.872323	1.000000	0.486049
ShapeFactor4	0.531714	0.486049	1.000000

the decision was taken to drop some features with correlations to area over .9,
the benefits of this decision should be a significant reduction of the computational complexity.
I chose to use major and minor axis length

0.9995381489007944

0.9194680458071666

	Area	Perimeter	MajorAxisLength	MinorAxisLength	AspectRatio	\
0	37500	728.191	275.840463	173.818266	1.586948	
1	37500	715.578	272.171813	175.668301	1.549351	
2	37511	718.350	267.039757	179.141937	1.490660	
3	37513	720.028	269.589608	177.510928	1.518721	
4	37514	725.847	269.881174	177.418223	1.521158	

	Eccentricity	ConvexArea	EquivDiameter	Extent	Solidity	roundness	\
0	0.776481	37944	218.509686	0.703406	0.988299	0.888690	
1	0.763818	37797	218.509686	0.786229	0.992142	0.920295	
2	0.741599	37868	218.541732	0.717365	0.990573	0.913474	
3	0.752626	37981	218.547558	0.780545	0.987678	0.909270	
4	0.753547	37920	218.550471	0.793309	0.989293	0.894773	

	Compactness	ShapeFactor1	ShapeFactor2	ShapeFactor3	ShapeFactor4
0	0.792160	0.007356	0.001787	0.627517	0.995836
1	0.802837	0.007258	0.001860	0.644548	0.998631
2	0.818387	0.007119	0.001970	0.669756	0.998379
3	0.810668	0.007187	0.001915	0.657182	0.998076
4	0.809803	0.007194	0.001908	0.655780	0.997545

This is the bean prediction of the sample data

[3 3 3 3 3]

