

# Notebook

November 9, 2025

```
[1]: import numpy as np
import pandas as pd

import seaborn as sns
from seaborn import heatmap

import scipy.stats as stats

import matplotlib.pyplot as plt

df = pd.read_csv('data/steel-plates-fault.csv', sep=',')
df.rename(columns={'V1': 'X_Minimum'}, inplace=True)
df.rename(columns={'V2': 'X_Maximum'}, inplace=True)
df.rename(columns={'V3': 'Y_Minimum'}, inplace=True)
df.rename(columns={'V4': 'Y_Maximum'}, inplace=True)
df.rename(columns={'V5': 'Pixels_Areas'}, inplace=True)
df.rename(columns={'V6': 'X_Perimeter'}, inplace=True)
df.rename(columns={'V7': 'Y_Perimeter'}, inplace=True)
df.rename(columns={'V8': 'Sum_of_Luminosity'}, inplace=True)
df.rename(columns={'V9': 'Minimum_of_Luminosity'}, inplace=True)
df.rename(columns={'V10': 'Maximum_of_Luminosity'}, inplace=True)
df.rename(columns={'V11': 'Length_of_Conveyer'}, inplace=True)
df.rename(columns={'V12': 'TypesOfSteel_A300'}, inplace=True)
df.rename(columns={'V13': 'TypesOfSteel_A400'}, inplace=True)
df.rename(columns={'V14': 'Steel_Plate_Thickness'}, inplace=True)
df.rename(columns={'V15': 'Edges_Index'}, inplace=True)
df.rename(columns={'V16': 'Empty_Index'}, inplace=True)
df.rename(columns={'V17': 'Square_Index'}, inplace=True)
df.rename(columns={'V18': 'Outside_X_Index'}, inplace=True)
df.rename(columns={'V19': 'Edges_X_Index'}, inplace=True)
df.rename(columns={'V20': 'Edges_Y_Index'}, inplace=True)
df.rename(columns={'V21': 'Outside_Global_Index'}, inplace=True)
df.rename(columns={'V22': 'LogOfAreas'}, inplace=True)
df.rename(columns={'V23': 'Log_X_Index'}, inplace=True)
df.rename(columns={'V24': 'Log_Y_Index'}, inplace=True)
df.rename(columns={'V25': 'Orientation_Index'}, inplace=True)
df.rename(columns={'V26': 'Luminosity_Index'}, inplace=True)
```

```

df.rename(columns={'V27': 'SigmoidOfAreas'}, inplace=True)
df.rename(columns={'V28': 'Pastry'}, inplace=True)
df.rename(columns={'V29': 'Z_Scratch'}, inplace=True)
df.rename(columns={'V30': 'K_Scratch'}, inplace=True)
df.rename(columns={'V31': 'Stains'}, inplace=True)
df.rename(columns={'V32': 'Dirtiness'}, inplace=True)
df.rename(columns={'V33': 'Bumps'}, inplace=True)
df.rename(columns={'Class': 'Class'}, inplace=True) # Other_Faults
df

```

	X_Minimum	X_Maximum	Y_Minimum	Y_Maximum	Pixels_Areas	X_Perimeter	\
0	42	50	270900	270944	267	17	
1	645	651	2538079	2538108	108	10	
2	829	835	1553913	1553931	71	8	
3	853	860	369370	369415	176	13	
4	1289	1306	498078	498335	2409	60	
...	...	...	...	...	...	...	\
1936	249	277	325780	325796	273	54	
1937	144	175	340581	340598	287	44	
1938	145	174	386779	386794	292	40	
1939	137	170	422497	422528	419	97	
1940	1261	1281	87951	87967	103	26	
	Y_Perimeter	Sum_of_Luminosity	Minimum_of_Luminosity	\			
0	44	24220	76				
1	30	11397	84				
2	19	7972	99				
3	45	18996	99				
4	260	246930	37				
...	...	...	...	...			
1936	22	35033	119				
1937	24	34599	112				
1938	22	37572	120				
1939	47	52715	117				
1940	22	11682	101				
	Maximum_of_Luminosity	...	Orientation_Index	Luminosity_Index	\		
0	108	...	0.8182	-0.2913			
1	123	...	0.7931	-0.1756			
2	125	...	0.6667	-0.1228			
3	126	...	0.8444	-0.1568			
4	126	...	0.9338	-0.1992			
...	...	...	...	...			
1936	141	...	-0.4286	0.0026			
1937	133	...	-0.4516	-0.0582			
1938	140	...	-0.4828	0.0052			
1939	140	...	-0.0606	-0.0171			

1940	133	...	-0.2000	-0.1139	
SigmoidOfAreas	Pastry	Z_Scratch	K_Scratch	Stains	Dirtiness
0	0.5822	1	0	0	0
1	0.2984	1	0	0	0
2	0.2150	1	0	0	0
3	0.5212	1	0	0	0
4	1.0000	1	0	0	0
...	...	...	...	...	...
1936	0.7254	0	0	0	0
1937	0.8173	0	0	0	0
1938	0.7079	0	0	0	0
1939	0.9919	0	0	0	0
1940	0.5296	0	0	0	0
Class					
0	1				
1	1				
2	1				
3	1				
4	1				
...	...				
1936	2				
1937	2				
1938	2				
1939	2				
1940	2				

[1941 rows x 34 columns]

```
[2]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1941 entries, 0 to 1940
Data columns (total 34 columns):
 #   Column           Non-Null Count Dtype  
 --- 
 0   X_Minimum        1941 non-null  int64  
 1   X_Maximum        1941 non-null  int64  
 2   Y_Minimum        1941 non-null  int64  
 3   Y_Maximum        1941 non-null  int64  
 4   Pixels_Areas     1941 non-null  int64  
 5   X_Perimeter      1941 non-null  int64  
 6   Y_Perimeter      1941 non-null  int64  
 7   Sum_of_Luminosity 1941 non-null  int64  
 8   Minimum_of_Luminosity 1941 non-null  int64  
 9   Maximum_of_Luminosity 1941 non-null  int64  
 10  Length_of_Conveyer 1941 non-null  int64
```

```

11 TypesOfSteel_A300      1941 non-null    int64
12 TypesOfSteel_A400      1941 non-null    int64
13 Steel_Plate_Thickness 1941 non-null    int64
14 Edges_Index            1941 non-null    float64
15 Empty_Index            1941 non-null    float64
16 Square_Index           1941 non-null    float64
17 Outside_X_Index        1941 non-null    float64
18 Edges_X_Index          1941 non-null    float64
19 Edges_Y_Index          1941 non-null    float64
20 Outside_Global_Index   1941 non-null    float64
21 LogOfAreas             1941 non-null    float64
22 Log_X_Index            1941 non-null    float64
23 Log_Y_Index            1941 non-null    float64
24 Orientation_Index      1941 non-null    float64
25 Luminosity_Index       1941 non-null    float64
26 SigmoidOfAreas         1941 non-null    float64
27 Pastry                 1941 non-null    int64
28 Z_Scratch               1941 non-null    int64
29 K_Scratch               1941 non-null    int64
30 Stains                  1941 non-null    int64
31 Dirtiness               1941 non-null    int64
32 Bumps                   1941 non-null    int64
33 Class                   1941 non-null    int64
dtypes: float64(13), int64(21)
memory usage: 515.7 KB

```

[3]: `print("Duplicates (rows):", df.duplicated().sum())`

Duplicates (rows): 0

[4]: `df.describe()`

	X_Minimum	X_Maximum	Y_Minimum	Y_Maximum	Pixels_Areas	\
count	1941.000000	1941.000000	1.941000e+03	1.941000e+03	1941.000000	
mean	571.136012	617.964451	1.650685e+06	1.650739e+06	1893.878413	
std	520.690671	497.627410	1.774578e+06	1.774590e+06	5168.459560	
min	0.000000	4.000000	6.712000e+03	6.724000e+03	2.000000	
25%	51.000000	192.000000	4.712530e+05	4.712810e+05	84.000000	
50%	435.000000	467.000000	1.204128e+06	1.204136e+06	174.000000	
75%	1053.000000	1072.000000	2.183073e+06	2.183084e+06	822.000000	
max	1705.000000	1713.000000	1.298766e+07	1.298769e+07	152655.000000	
	X_Perimeter	Y_Perimeter	Sum_of_Luminosity	Minimum_of_Luminosity	\	
count	1941.000000	1941.000000	1.941000e+03	1941.000000		
mean	111.855229	82.965997	2.063121e+05	84.548686		
std	301.209187	426.482879	5.122936e+05	32.134276		
min	2.000000	1.000000	2.500000e+02	0.000000		
25%	15.000000	13.000000	9.522000e+03	63.000000		

50%	26.000000	25.000000	1.920200e+04	90.000000
75%	84.000000	83.000000	8.301100e+04	106.000000
max	10449.000000	18152.000000	1.159141e+07	203.000000

	Maximum_of_Luminosity	...	Orientation_Index	Luminosity_Index	\
count	1941.000000	...	1941.000000	1941.000000	
mean	130.193715	...	0.083288	-0.131305	
std	18.690992	...	0.500868	0.148767	
min	37.000000	...	-0.991000	-0.998900	
25%	124.000000	...	-0.333300	-0.195000	
50%	127.000000	...	0.095200	-0.133000	
75%	140.000000	...	0.511600	-0.066600	
max	253.000000	...	0.991700	0.642100	

	SigmoidOfAreas	Pastry	Z_Scratch	K_Scratch	Stains	\
count	1941.000000	1941.000000	1941.000000	1941.000000	1941.000000	
mean	0.585420	0.081401	0.097888	0.201443	0.037094	
std	0.339452	0.273521	0.297239	0.401181	0.189042	
min	0.119000	0.000000	0.000000	0.000000	0.000000	
25%	0.248200	0.000000	0.000000	0.000000	0.000000	
50%	0.506300	0.000000	0.000000	0.000000	0.000000	
75%	0.999800	0.000000	0.000000	0.000000	0.000000	
max	1.000000	1.000000	1.000000	1.000000	1.000000	

	Dirtiness	Bumps	Class
count	1941.000000	1941.000000	1941.000000
mean	0.028336	0.207110	1.346728
std	0.165973	0.405339	0.476051
min	0.000000	0.000000	1.000000
25%	0.000000	0.000000	1.000000
50%	0.000000	0.000000	1.000000
75%	0.000000	0.000000	2.000000
max	1.000000	1.000000	2.000000

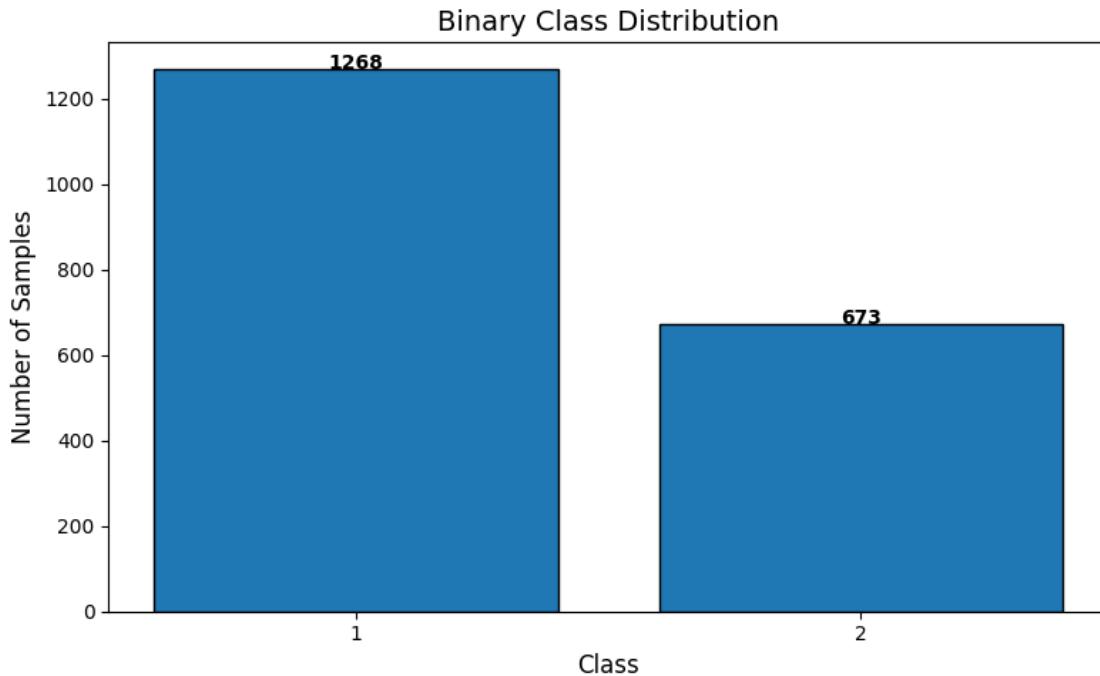
[8 rows x 34 columns]

```
[5]: import matplotlib.pyplot as plt

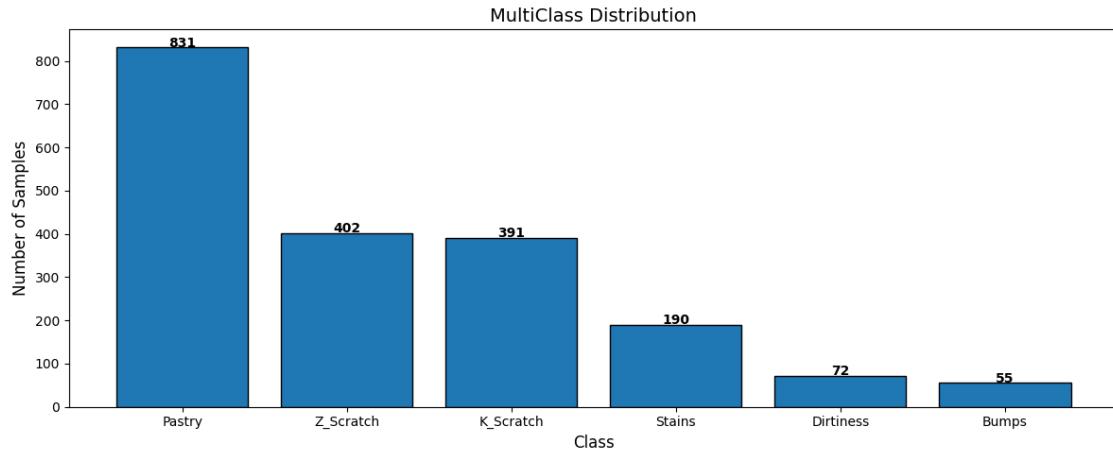
counts = df["Class"].value_counts()

plt.figure(figsize=(8, 5))
plt.bar(counts.index.astype(str), counts.values, edgecolor="black")
plt.title("Binary Class Distribution", fontsize=14)
plt.xlabel("Class", fontsize=12)
plt.ylabel("Number of Samples", fontsize=12)
for i, v in enumerate(counts.values):
    plt.text(i, v + 0.5, str(v), ha='center', fontweight='bold')
```

```
plt.tight_layout()  
plt.show()
```



```
[6]: import matplotlib.pyplot as plt  
  
class_names = "Pastry Z_Scratch K_Scratch Stains Dirtiness Bumps".split(" ")  
counts = df[class_names].dot([i for i in range(6)]).value_counts()  
  
plt.figure(figsize=(12, 5))  
plt.bar(class_names, counts.values, edgecolor="black")  
plt.title("MultiClass Distribution", fontsize=14)  
plt.xlabel("Class", fontsize=12)  
plt.ylabel("Number of Samples", fontsize=12)  
  
for i, v in enumerate(counts.values):  
    plt.text(i, v + 0.5, str(v), ha='center', fontweight='bold')  
plt.tight_layout()  
plt.show()
```



```
[7]: "Pastry Z_Scratch K_Scratch Stains Dirtiness Bumps".split(" ")
```

```
[7]: ['Pastry', 'Z_Scratch', 'K_Scratch', 'Stains', 'Dirtiness', 'Bumps']
```

```
[8]: num = df.select_dtypes(include='number')
corr_matrix = num.corr(method='pearson')
corr_matrix
```

	X_Minimum	X_Maximum	Y_Minimum	Y_Maximum	\
X_Minimum	1.000000	0.988314	0.041821	0.041807	
X_Maximum	0.988314	1.000000	0.052147	0.052135	
Y_Minimum	0.041821	0.052147	1.000000	1.000000	
Y_Maximum	0.041807	0.052135	1.000000	1.000000	
Pixels_Areas	-0.307322	-0.225399	0.017670	0.017840	
X_Perimeter	-0.258937	-0.186326	0.023843	0.024038	
Y_Perimeter	-0.118757	-0.090138	0.024150	0.024380	
Sum_of_Luminosity	-0.339045	-0.247052	0.007362	0.007499	
Minimum_of_Luminosity	0.237637	0.168649	-0.065703	-0.065733	
Maximum_of_Luminosity	-0.075554	-0.062392	-0.067785	-0.067776	
Length_of_Conveyer	0.316662	0.299390	-0.049211	-0.049219	
TypesOfSteel_A300	0.144319	0.112009	0.075164	0.075151	
TypesOfSteel_A400	-0.144319	-0.112009	-0.075164	-0.075151	
Steel_Plate_Thickness	0.136625	0.106119	-0.207640	-0.207644	
Edges_Index	0.278075	0.242846	0.021314	0.021300	
Empty_Index	-0.198461	-0.152680	-0.043117	-0.043085	
Square_Index	0.063658	0.048575	-0.006135	-0.006152	
Outside_X_Index	-0.361160	-0.214930	0.054165	0.054185	
Edges_X_Index	0.154778	0.149259	0.066085	0.066051	
Edges_Y_Index	0.367907	0.271915	-0.036543	-0.036549	
Outside_Global_Index	0.147282	0.099253	-0.062911	-0.062901	
LogOfAreas	-0.428553	-0.332169	0.044952	0.044994	

Log_X_Index	-0.437944	-0.324012	0.070406	0.070432
Log_Y_Index	-0.326851	-0.265990	-0.008442	-0.008382
Orientation_Index	0.178585	0.115019	-0.086497	-0.086480
Luminosity_Index	-0.031578	-0.038996	-0.090654	-0.090666
SigmoidOfAreas	-0.355251	-0.286736	0.025257	0.025284
Pastry	0.134956	0.119814	0.036488	0.036488
Z_Scratch	-0.228960	-0.258178	-0.063327	-0.063329
K_Scratch	-0.419264	-0.336084	-0.000420	-0.000397
Stains	0.073740	0.061471	-0.066601	-0.066606
Dirtiness	0.103924	0.096523	0.064262	0.064262
Bumps	0.221296	0.201704	0.126121	0.126110
Class	0.164804	0.145783	-0.084415	-0.084422

	Pixels_Areas	X_Perimeter	Y_Perimeter	\
X_Minimum	-0.307322	-0.258937	-0.118757	
X_Maximum	-0.225399	-0.186326	-0.090138	
Y_Minimum	0.017670	0.023843	0.024150	
Y_Maximum	0.017840	0.024038	0.024380	
Pixels_Areas	1.000000	0.966644	0.827199	
X_Perimeter	0.966644	1.000000	0.912436	
Y_Perimeter	0.827199	0.912436	1.000000	
Sum_of_Luminosity	0.978952	0.912956	0.704876	
Minimum_of_Luminosity	-0.497204	-0.400427	-0.213758	
Maximum_of_Luminosity	0.110063	0.111363	0.061809	
Length_of_Conveyer	-0.155853	-0.134240	-0.063825	
TypesOfSteel_A300	-0.235591	-0.189250	-0.095154	
TypesOfSteel_A400	0.235591	0.189250	0.095154	
Steel_Plate_Thickness	-0.183735	-0.147712	-0.058889	
Edges_Index	-0.275289	-0.227590	-0.111240	
Empty_Index	0.272808	0.306348	0.188825	
Square_Index	0.017865	0.004507	-0.047511	
Outside_X_Index	0.588606	0.517098	0.209160	
Edges_X_Index	-0.294673	-0.293039	-0.195162	
Edges_Y_Index	-0.463571	-0.412100	-0.136723	
Outside_Global_Index	-0.109655	-0.079106	0.013438	
LogOfAreas	0.650234	0.563036	0.294040	
Log_X_Index	0.603072	0.524716	0.228485	
Log_Y_Index	0.578342	0.523472	0.344378	
Orientation_Index	-0.137604	-0.101731	0.031381	
Luminosity_Index	-0.043449	-0.032617	-0.047778	
SigmoidOfAreas	0.422947	0.380605	0.191772	
Pastry	-0.076752	-0.075418	-0.017616	
Z_Scratch	-0.088440	-0.060582	-0.025721	
K_Scratch	0.556846	0.455003	0.203063	
Stains	-0.071182	-0.067547	-0.035743	
Dirtiness	-0.050578	-0.037820	-0.010058	
Bumps	-0.163739	-0.140197	-0.070989	

Class	-0.184632	-0.142903	-0.066801
	Sum_of_Luminosity	Minimum_of_Luminosity	\
X_Minimum	-0.339045	0.237637	
X_Maximum	-0.247052	0.168649	
Y_Minimum	0.007362	-0.065703	
Y_Maximum	0.007499	-0.065733	
Pixels_Areas	0.978952	-0.497204	
X_Perimeter	0.912956	-0.400427	
Y_Perimeter	0.704876	-0.213758	
Sum_of_Luminosity	1.000000	-0.540566	
Minimum_of_Luminosity	-0.540566	1.000000	
Maximum_of_Luminosity	0.136515	0.429605	
Length_of_Conveyer	-0.169331	-0.023579	
TypesOfSteel_A300	-0.263632	0.042048	
TypesOfSteel_A400	0.263632	-0.042048	
Steel_Plate_Thickness	-0.204812	0.103393	
Edges_Index	-0.301452	0.358915	
Empty_Index	0.293691	-0.044111	
Square_Index	0.049607	0.066748	
Outside_X_Index	0.658339	-0.487574	
Edges_X_Index	-0.327728	0.252256	
Edges_Y_Index	-0.529745	0.316610	
Outside_Global_Index	-0.121090	0.035462	
LogOfAreas	0.712128	-0.678762	
Log_X_Index	0.667736	-0.567655	
Log_Y_Index	0.618795	-0.588208	
Orientation_Index	-0.158483	0.057123	
Luminosity_Index	-0.014067	0.669534	
SigmoidOfAreas	0.464248	-0.514797	
Pastry	-0.084307	-0.074697	
Z_Scratch	-0.099592	0.049905	
K_Scratch	0.616950	-0.461000	
Stains	-0.078111	0.183327	
Dirtiness	-0.055272	0.092765	
Bumps	-0.179831	0.078690	
Class	-0.205890	0.228112	
	Maximum_of_Luminosity	...	Orientation_Index \
X_Minimum	-0.075554	...	0.178585
X_Maximum	-0.062392	...	0.115019
Y_Minimum	-0.067785	...	-0.086497
Y_Maximum	-0.067776	...	-0.086480
Pixels_Areas	0.110063	...	-0.137604
X_Perimeter	0.111363	...	-0.101731
Y_Perimeter	0.061809	...	0.031381
Sum_of_Luminosity	0.136515	...	-0.158483

Minimum_of_Luminosity	0.429605	...	0.057123	
Maximum_of_Luminosity	1.000000	...	-0.169747	
Length_of_Conveyer	-0.098009	...	0.120715	
TypesOfSteel_A300	-0.216339	...	0.010630	
TypesOfSteel_A400	0.216339	...	-0.010630	
Steel_Plate_Thickness	-0.128397	...	0.274097	
Edges_Index	0.149675	...	0.020548	
Empty_Index	0.031425	...	-0.139420	
Square_Index	0.065517	...	-0.162034	
Outside_X_Index	0.099300	...	-0.440358	
Edges_X_Index	0.093522	...	-0.550302	
Edges_Y_Index	-0.167441	...	0.658049	
Outside_Global_Index	-0.124039	...	0.862670	
LogOfAreas	0.007672	...	-0.123898	
Log_X_Index	0.092823	...	-0.536629	
Log_Y_Index	-0.069522	...	0.316792	
Orientation_Index	-0.169747	...	1.000000	
Luminosity_Index	0.870160	...	-0.153464	
SigmoidOfAreas	-0.039651	...	-0.023978	
Pastry	-0.058742	...	0.329385	
Z_Scratch	-0.189441	...	0.117501	
K_Scratch	0.185897	...	-0.384696	
Stains	0.090456	...	-0.136713	
Dirtiness	0.022157	...	0.174628	
Bumps	-0.047549	...	0.013614	
Class	-0.007784	...	0.043389	
X_Minimum	-0.031578	-0.355251	0.134956	-0.228960
X_Maximum	-0.038996	-0.286736	0.119814	-0.258178
Y_Minimum	-0.090654	0.025257	0.036488	-0.063327
Y_Maximum	-0.090666	0.025284	0.036488	-0.063329
Pixels_Areas	-0.043449	0.422947	-0.076752	-0.088440
X_Perimeter	-0.032617	0.380605	-0.075418	-0.060582
Y_Perimeter	-0.047778	0.191772	-0.017616	-0.025721
Sum_of_Luminosity	-0.014067	0.464248	-0.084307	-0.099592
Minimum_of_Luminosity	0.669534	-0.514797	-0.074697	0.049905
Maximum_of_Luminosity	0.870160	-0.039651	-0.058742	-0.189441
Length_of_Conveyer	-0.149769	-0.197543	0.196209	-0.230856
TypesOfSteel_A300	-0.252818	-0.308910	-0.054792	0.339488
TypesOfSteel_A400	0.252818	0.308910	0.054792	-0.339488
Steel_Plate_Thickness	-0.116499	-0.085159	0.076579	-0.024876
Edges_Index	0.207516	-0.330006	-0.029915	-0.152730
Empty_Index	0.061608	0.481738	-0.184993	0.129672
Square_Index	0.111977	-0.292251	-0.242923	-0.005543
Outside_X_Index	-0.035721	0.518910	-0.132081	-0.109354
Edges_X_Index	0.126460	-0.558426	-0.122428	-0.141291

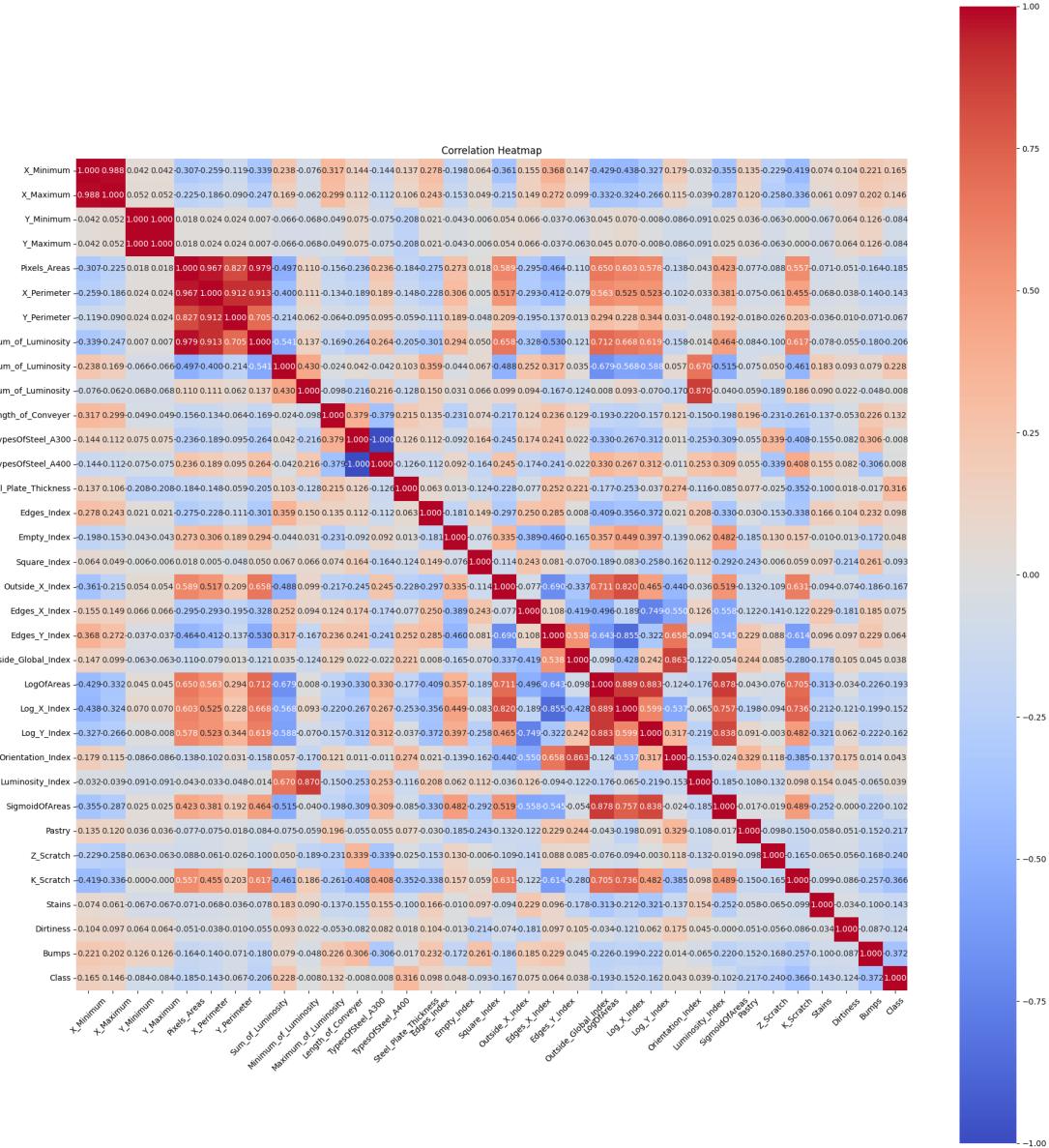
Edges_Y_Index	-0.094368	-0.545393	0.229432	0.088045
Outside_Global_Index	-0.122321	-0.053770	0.244320	0.084885
LogOfAreas	-0.175879	0.877768	-0.042526	-0.075564
Log_X_Index	-0.064923	0.757343	-0.198250	-0.094426
Log_Y_Index	-0.219110	0.838188	0.091130	-0.003170
Orientation_Index	-0.153464	-0.023978	0.329385	0.117501
Luminosity_Index	1.000000	-0.184840	-0.108018	-0.131623
SigmoidOfAreas	-0.184840	1.000000	-0.017422	-0.019397
Pastry	-0.108018	-0.017422	1.000000	-0.098059
Z_Scratch	-0.131623	-0.019397	-0.098059	1.000000
K_Scratch	0.098459	0.488878	-0.149512	-0.165446
Stains	0.154319	-0.252488	-0.058427	-0.064654
Dirtiness	0.045295	-0.000462	-0.050835	-0.056253
Bumps	-0.064744	-0.220091	-0.152141	-0.168356
Class	0.039328	-0.102046	-0.216871	-0.239984

	K_Scratch	Stains	Dirtiness	Bumps	Class
X_Minimum	-0.419264	0.073740	0.103924	0.221296	0.164804
X_Maximum	-0.336084	0.061471	0.096523	0.201704	0.145783
Y_Minimum	-0.000420	-0.066601	0.064262	0.126121	-0.084415
Y_Maximum	-0.000397	-0.066606	0.064262	0.126110	-0.084422
Pixels_Areas	0.556846	-0.071182	-0.050578	-0.163739	-0.184632
X_Perimeter	0.455003	-0.067547	-0.037820	-0.140197	-0.142903
Y_Perimeter	0.203063	-0.035743	-0.010058	-0.070989	-0.066801
Sum_of_Luminosity	0.616950	-0.078111	-0.055272	-0.179831	-0.205890
Minimum_of_Luminosity	-0.461000	0.183327	0.092765	0.078690	0.228112
Maximum_of_Luminosity	0.185897	0.090456	0.022157	-0.047549	-0.007784
Length_of_Conveyer	-0.261071	-0.136839	-0.052603	0.225504	0.132091
TypesOfSteel_A300	-0.407730	-0.154796	-0.082489	0.306385	-0.007530
TypesOfSteel_A400	0.407730	0.154796	0.082489	-0.306385	0.007530
Steel_Plate_Thickness	-0.351654	-0.099945	0.017727	-0.016773	0.315671
Edges_Index	-0.337701	0.165732	0.103517	0.232000	0.097698
Empty_Index	0.156711	-0.010243	-0.012514	-0.172147	0.048267
Square_Index	0.059175	0.097310	-0.214369	0.261385	-0.093296
Outside_X_Index	0.631370	-0.093722	-0.074105	-0.185729	-0.166709
Edges_X_Index	-0.121541	0.229406	-0.180588	0.185481	0.074921
Edges_Y_Index	-0.614341	0.096336	0.096862	0.229231	0.063718
Outside_Global_Index	-0.279992	-0.177802	0.105173	0.044964	0.038231
LogOfAreas	0.704531	-0.312690	-0.034345	-0.226490	-0.193121
Log_X_Index	0.735860	-0.212375	-0.120561	-0.198690	-0.151720
Log_Y_Index	0.481853	-0.321185	0.062316	-0.221813	-0.161769
Orientation_Index	-0.384696	-0.136713	0.174628	0.013614	0.043389
Luminosity_Index	0.098459	0.154319	0.045295	-0.064744	0.039328
SigmoidOfAreas	0.488878	-0.252488	-0.000462	-0.220091	-0.102046
Pastry	-0.149512	-0.058427	-0.050835	-0.152141	-0.216871
Z_Scratch	-0.165446	-0.064654	-0.056253	-0.168356	-0.239984
K_Scratch	1.000000	-0.098579	-0.085770	-0.256694	-0.365907

```
Stains           -0.098579  1.000000  -0.033518 -0.100313 -0.142991
Dirtiness        -0.085770 -0.033518   1.000000 -0.087278 -0.124411
Bumps            -0.256694 -0.100313  -0.087278  1.000000 -0.372342
Class            -0.365907 -0.142991  -0.124411 -0.372342  1.000000
```

[34 rows x 34 columns]

```
[9]: plt.figure(figsize=(20, 20))
sns.heatmap(corr_matrix, cmap='coolwarm', fmt=".3f", vmin=-1, vmax=1, center=0,
             annot=True, square=True)
plt.title("Correlation Heatmap")
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```



```
[10]: plt.figure()
nitems = len(df.columns)
ncols = 5
nrows = (nitems + ncols - 1) // ncols

fig, axes = plt.subplots(nrows, ncols, figsize=(15, 5 * nrows))
axes = axes.flatten()

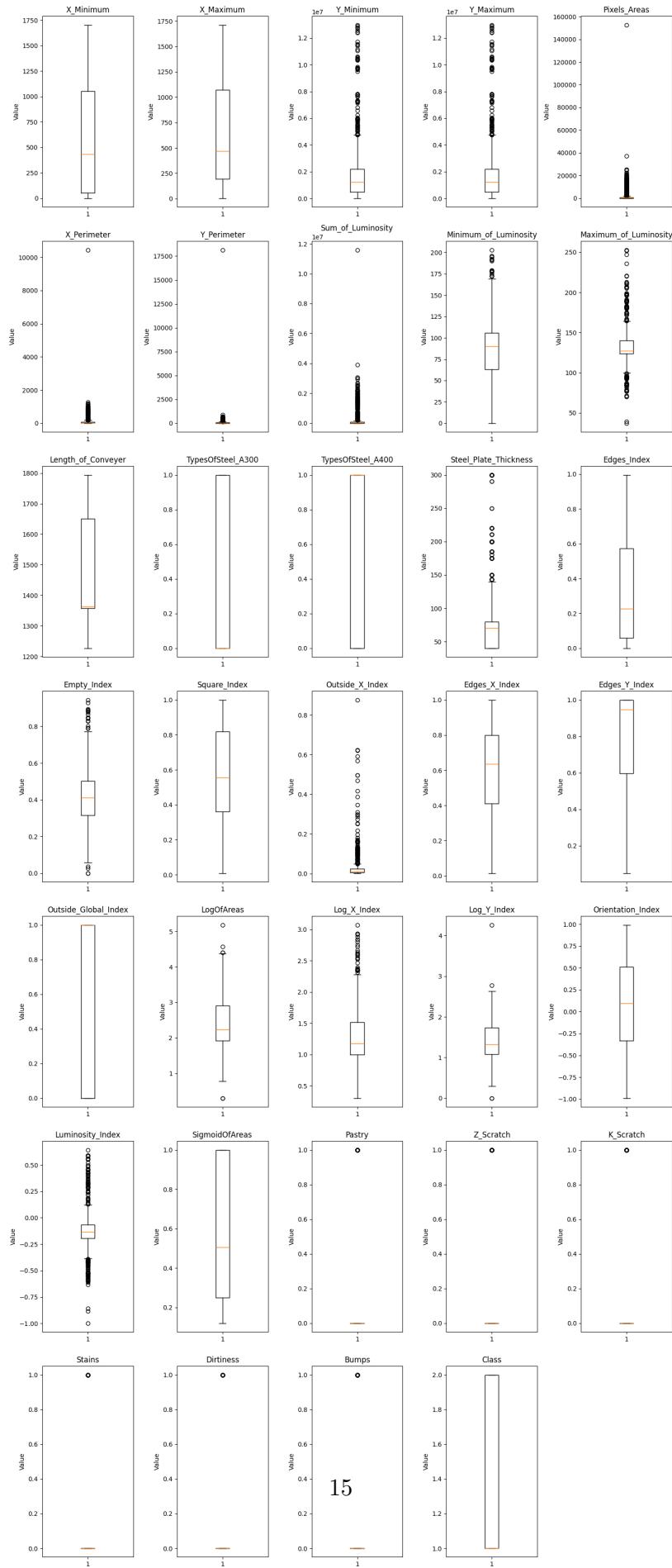
for i, col in enumerate(df.columns):
    axes[i].boxplot(df[col].dropna())
    axes[i].set_title(col)
```

```
axes[i].set_ylabel("Value")

for i in range(nitems, len(axes)):
    fig.delaxes(axes[i])

plt.tight_layout()
plt.show()
```

<Figure size 640x480 with 0 Axes>



## 1 Z-Score outlier detection

```
[11]: # All Columns except the ones we're predicting ('Pastry', 'Z_Scratch', ↴'K_Scratch', 'Stains', 'Dirtiness', 'Bumps', 'Other_Faults_(Class)')
columns_to_check_outliers = [
    'X_Minimum', 'X_Maximum', 'Y_Minimum', 'Y_Maximum', 'Pixels_Areas', ↴
    ↴'X_Perimeter',
    'Y_Perimeter', 'Sum_of_Luminosity', 'Minimum_of_Luminosity', ↴
    ↴'Maximum_of_Luminosity',
    'Length_of_Conveyer', 'TypesOfSteel_A300', 'TypesOfSteel_A400', ↴
    ↴'Steel_Plate_Thickness',
    'Edges_Index', 'Empty_Index', 'Square_Index', 'Outside_X_Index', ↴
    ↴'Edges_X_Index',
    'Edges_Y_Index', 'Outside_Global_Index', 'LogOfAreas', 'Log_X_Index', ↴
    ↴'Log_Y_Index',
    'Orientation_Index', 'Luminosity_Index', 'SigmoidOfAreas'
]
# Ignoring certain columns without outliers - as seen in the boxplots - to use ↴
# a more aggressive outlier detection on the others
columns_to_ignore = [
    'X_Minimum', 'X_Maximum', 'Length_of_Conveyer', 'TypesOfSteel_A300', ↴
    ↴'TypesOfSteel_A400',
    'Edges_Index', 'Square_Index', 'Edges_X_Index', 'Edges_Y_Index', ↴
    ↴'Outside_Global_Index',
    'Orientation_Index', 'SigmoidOfAreas'
]
columns_to_check_outliers = [col for col in columns_to_check_outliers if col ↴
    ↴not in columns_to_ignore]
print(columns_to_check_outliers)
z_score = stats.zscore(df[columns_to_check_outliers])
df_clean = df.copy()
df_clean = df_clean[(abs(z_score) < 4).all(axis=1)]
```

['Y\_Minimum', 'Y\_Maximum', 'Pixels\_Areas', 'X\_Perimeter', 'Y\_Perimeter',  
 'Sum\_of\_Luminosity', 'Minimum\_of\_Luminosity', 'Maximum\_of\_Luminosity',  
 'Steel\_Plate\_Thickness', 'Empty\_Index', 'Outside\_X\_Index', 'LogOfAreas',  
 'Log\_X\_Index', 'Log\_Y\_Index', 'Luminosity\_Index']

```
[12]: plt.figure()
nitems = len(df_clean.columns)
ncols = 5
nrows = (nitems + ncols - 1) // ncols

fig, axes = plt.subplots(nrows, ncols, figsize=(15, 5 * nrows))
```

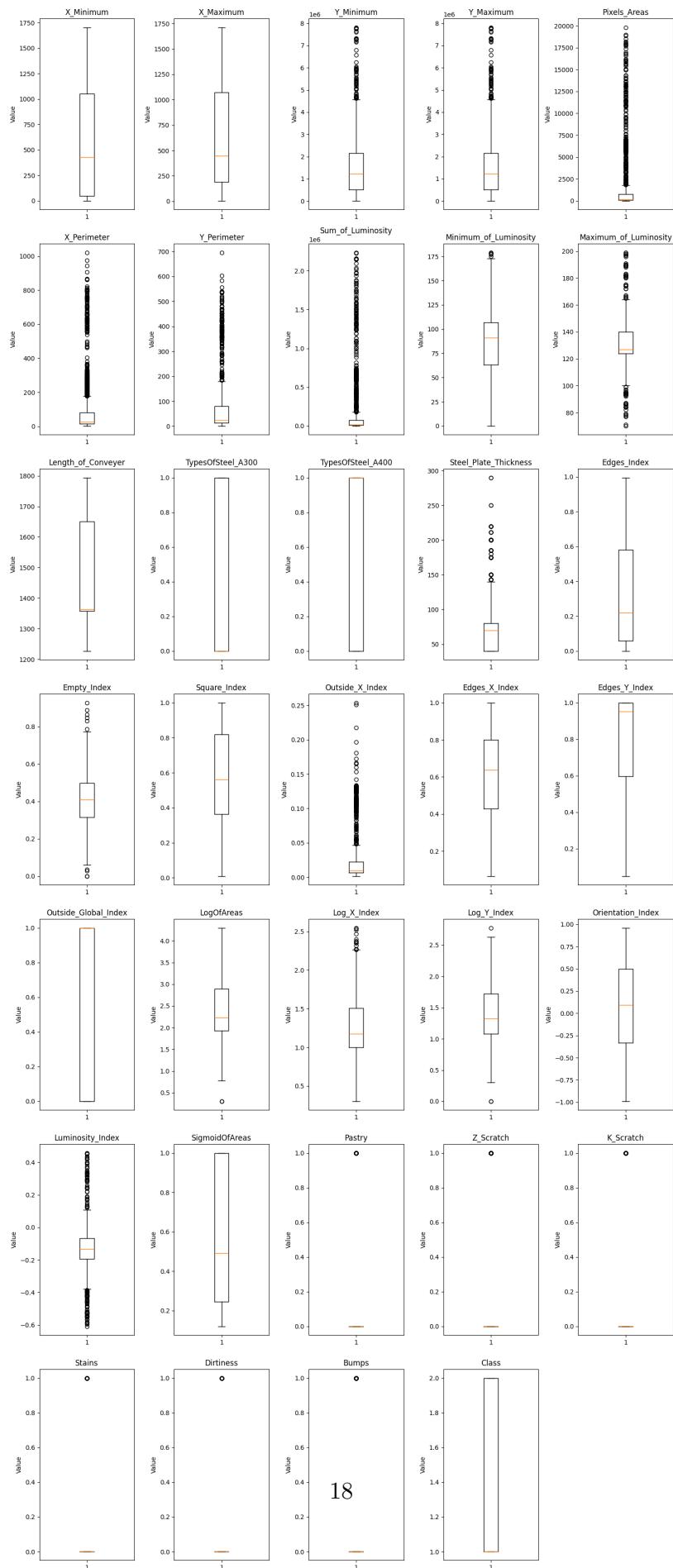
```
axes = axes.flatten()

for i, col in enumerate(df_clean.columns):
    axes[i].boxplot(df_clean[col].dropna())
    axes[i].set_title(col)
    axes[i].set_ylabel("Value")

for i in range(nitems, len(axes)):
    fig.delaxes(axes[i])

plt.tight_layout()
plt.show()
```

<Figure size 640x480 with 0 Axes>



```
[13]: df_clean.describe()
```

	X_Minimum	X_Maximum	Y_Minimum	Y_Maximum	Pixels_Areas	\
count	1822.000000	1822.000000	1.822000e+03	1.822000e+03	1822.000000	
mean	568.457190	610.583425	1.545173e+06	1.545216e+06	1659.091109	
std	521.583285	499.458726	1.351809e+06	1.351807e+06	3407.213291	
min	0.000000	4.000000	6.712000e+03	6.724000e+03	2.000000	
25%	50.000000	190.000000	5.144658e+05	5.146172e+05	84.000000	
50%	430.500000	446.000000	1.230772e+06	1.230788e+06	170.000000	
75%	1051.000000	1071.000000	2.156948e+06	2.156966e+06	780.000000	
max	1705.000000	1713.000000	7.818756e+06	7.818807e+06	19818.000000	
	X_Perimeter	Y_Perimeter	Sum_of_Luminosity	Minimum_of_Luminosity	\	
count	1822.000000	1822.000000	1.822000e+03	1822.000000		
mean	98.555434	68.99067	1.821127e+05	84.665203		
std	168.790145	105.22777	3.857515e+05	30.904734		
min	2.000000	1.000000	2.500000e+02	0.000000		
25%	15.000000	13.000000	9.453500e+03	63.000000		
50%	26.000000	24.000000	1.868450e+04	91.000000		
75%	80.000000	80.000000	7.780275e+04	107.000000		
max	1021.000000	696.000000	2.236201e+06	179.000000		
	Maximum_of_Luminosity	...	Orientation_Index	Luminosity_Index	\	
count	1822.000000	...	1822.000000	1822.000000		
mean	129.660263	...	0.082325	-0.132847		
std	16.070491	...	0.499101	0.133654		
min	70.000000	...	-0.991000	-0.609600		
25%	124.000000	...	-0.333300	-0.193450		
50%	127.000000	...	0.090900	-0.132550		
75%	140.000000	...	0.500000	-0.067950		
max	199.000000	...	0.960700	0.457300		
	SigmoidOfAreas	Pastry	Z_Scratch	K_Scratch	Stains	\
count	1822.000000	1822.000000	1822.000000	1822.000000	1822.000000	
mean	0.579185	0.083425	0.104281	0.197585	0.039517	
std	0.338397	0.276599	0.305709	0.398287	0.194875	
min	0.119000	0.000000	0.000000	0.000000	0.000000	
25%	0.244500	0.000000	0.000000	0.000000	0.000000	
50%	0.491300	0.000000	0.000000	0.000000	0.000000	
75%	0.999700	0.000000	0.000000	0.000000	0.000000	
max	1.000000	1.000000	1.000000	1.000000	1.000000	
	Dirtiness	Bumps	Class			
count	1822.000000	1822.000000	1822.000000			

```

mean      0.030187    0.215697    1.329308
std       0.171147    0.411418    0.470091
min       0.000000    0.000000    1.000000
25%      0.000000    0.000000    1.000000
50%      0.000000    0.000000    1.000000
75%      0.000000    0.000000    2.000000
max       1.000000    1.000000    2.000000

[8 rows x 34 columns]

```

### 1.0.1 Zhodnotenie Z-Score outlier detection

Napriek tomu, že Z-Score efektívne odstránil väčšinu outlierov, prišli sme o dátu, ktoré nám pomôžu predikovať Pastry, Z Scratch, Stains a Dirtiness a teda Z-Score neoptimálny v tomto prípade alebo v aktuálnej konfigurácii.

## 2 Modifikovaná 5% 95% Metóda

```
[14]: # All Columns except the ones we're predicting ('Pastry', 'Z_Scratch', ↴
    ↴ 'K_Scratch', 'Stains', 'Dirtiness', 'Bumps', 'Other_Faults_(Class)')
columns_to_check_outliers = [
    'X_Minimum', 'X_Maximum', 'Y_Minimum', 'Y_Maximum', 'Pixels_Areas', ↴
    ↴ 'X_Perimeter',
    'Y_Perimeter', 'Sum_of_Luminosity', 'Minimum_of_Luminosity', ↴
    ↴ 'Maximum_of_Luminosity',
    'TypesOfSteel_A300', 'TypesOfSteel_A400', 'Steel_Plate_Thickness', ↴
    ↴ 'Edges_Index',
    'Empty_Index', 'Square_Index', 'Outside_X_Index', 'Edges_X_Index', ↴
    ↴ 'Edges_Y_Index',
    'Outside_Global_Index', 'LogOfAreas', 'Log_X_Index', 'Log_Y_Index', ↴
    ↴ 'Orientation_Index',
    'Luminosity_Index', 'SigmoidOfAreas'
]
# Ignoring certain columns without outliers - as seen in the boxplots - to use ↴
# a more aggressive outlier detection on the others
columns_to_ignore = [
    'X_Minimum', 'X_Maximum', 'Length_of_Conveyer', 'TypesOfSteel_A300', ↴
    ↴ 'TypesOfSteel_A400',
    'Edges_Index', 'Square_Index', 'Edges_X_Index', 'Edges_Y_Index', ↴
    ↴ 'Outside_Global_Index',
    'Orientation_Index', 'SigmoidOfAreas',
    # Ignoring columns that were well handled by Z-Score already
    'Y_Minimum', 'Y_Maximum', 'Minimum_of_Luminosity', 'Maximum_of_Luminosity', ↴
    ↴ 'Steel_Plate_Thickness', 'Empty_Index',
    'Log_X_Index', 'Log_Y_Index', 'Luminosity_Index'
]
```

```

columns_to_check_outliers = [col for col in columns_to_check_outliers if col not in columns_to_ignore]
print(columns_to_check_outliers)

shaving_ranges = [
    (0.0, 0.98), (0.0, 0.98), (0.0, 0.98), (0.0, 0.98), (0.0, 0.98), (0.0, 0.98)
]

for idx, col in enumerate(columns_to_check_outliers):
    lower = 0
    higher = 1
    df_CO_lower = df_clean[col].quantile(shaving_ranges[idx][lower])
    df_CO_upper = df_clean[col].quantile(shaving_ranges[idx][higher])

    df_clean[col] = df[col].where(
        (df_clean[col] >= df_CO_lower) & (df_clean[col] <= df_CO_upper)
    )
    df_clean.dropna(inplace = True)

```

['Pixels\_Areas', 'X\_Perimeter', 'Y\_Perimeter', 'Sum\_of\_Luminosity',  
'Outside\_X\_Index', 'Log0fAreas']

```

[15]: plt.figure()
nitems = len(df_clean.columns)
ncols = 5
nrows = (nitems + ncols - 1) // ncols

fig, axes = plt.subplots(nrows, ncols, figsize=(15, 5 * nrows))
axes = axes.flatten()

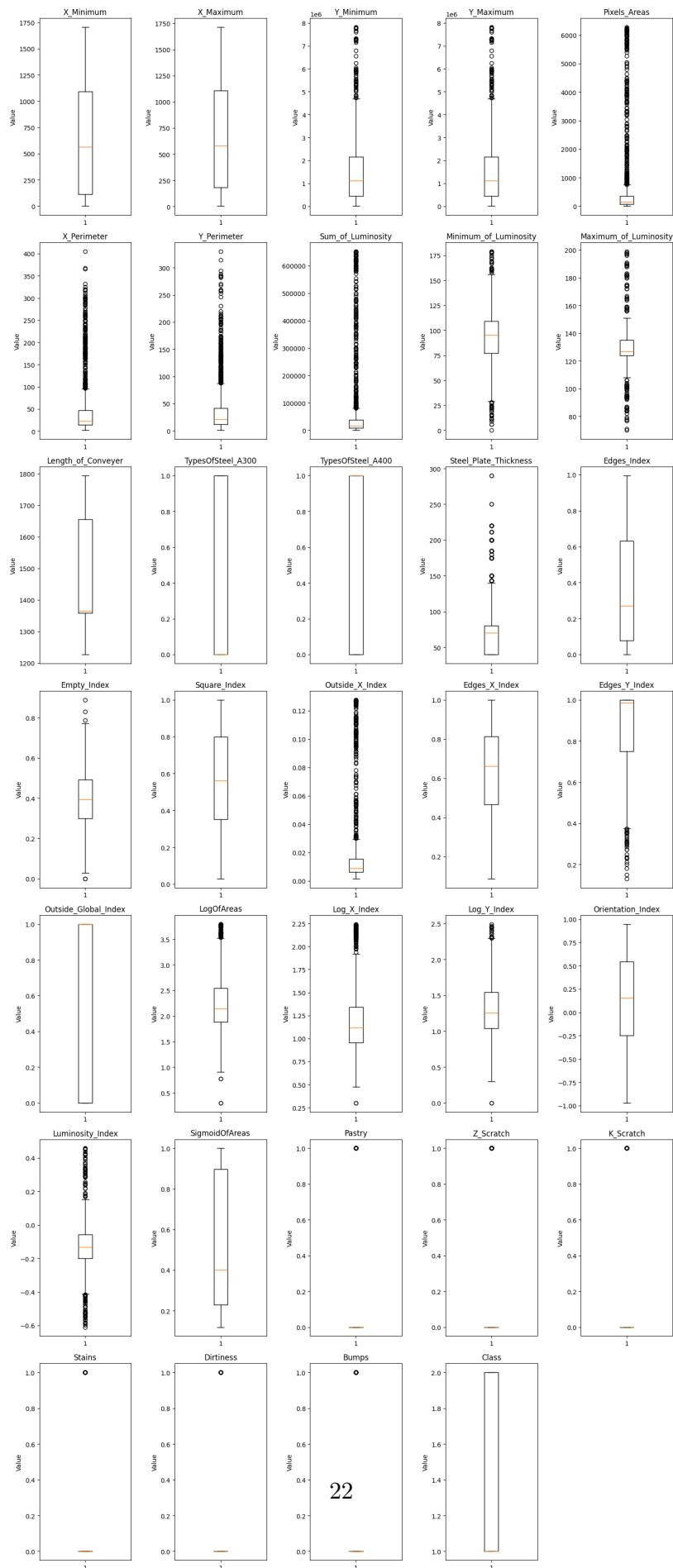
for i, col in enumerate(df_clean.columns):
    axes[i].boxplot(df_clean[col].dropna())
    axes[i].set_title(col)
    axes[i].set_ylabel("Value")

for i in range(nitems, len(axes)):
    fig.delaxes(axes[i])

plt.tight_layout()
plt.show()

```

<Figure size 640x480 with 0 Axes>



```
[16]: df_clean.describe()
```

```
[16]:      X_Minimum    X_Maximum    Y_Minimum    Y_Maximum  Pixels_Areas \
count  1613.000000  1613.000000  1.613000e+03  1.613000e+03  1613.000000
mean   629.401116  655.522009  1.513393e+06  1.513424e+06  632.513329
std    517.427471  507.119546  1.403003e+06  1.403001e+06  1334.099392
min    0.000000    4.000000   6.712000e+03  6.724000e+03  2.000000
25%   114.000000  181.000000  4.479430e+05  4.479920e+05  77.000000
50%   563.000000  583.000000  1.116878e+06  1.116892e+06  141.000000
75%   1090.000000 1106.000000  2.150529e+06  2.150756e+06  349.000000
max   1705.000000 1713.000000  7.818756e+06  7.818807e+06  6277.000000

      X_Perimeter  Y_Perimeter  Sum_of_Luminosity  Minimum_of_Luminosity \
count  1613.000000  1613.000000  1613.000000  1613.000000
mean   48.324241   39.787353   66874.983881   90.650961
std    64.058738   47.378852   139509.821414   27.155015
min    2.000000    1.000000    250.000000    0.000000
25%   14.000000   12.000000   8602.000000   77.000000
50%   22.000000   21.000000   16182.000000  95.000000
75%   47.000000   42.000000   37460.000000  109.000000
max   405.000000  330.000000  652005.000000  179.000000

      Maximum_of_Luminosity ...  Orientation_Index  Luminosity_Index \
count  1613.000000 ...  1613.000000  1613.000000
mean   129.055797 ...  0.121622  -0.131680
std    16.596791 ...  0.495092  0.140006
min    70.000000 ...  -0.970600  -0.609600
25%   124.000000 ...  -0.250000  -0.199500
50%   127.000000 ...  0.153900  -0.132600
75%   135.000000 ...  0.545400  -0.057500
max   199.000000 ...  0.946700  0.457300

      SigmoidOfAreas  Pastry  Z_Scratch  K_Scratch  Stains \
count  1613.000000  1613.000000  1613.000000  1613.000000  1613.000000
mean   0.524665   0.091754   0.115313   0.106634   0.044637
std    0.321606   0.288769   0.319498   0.308743   0.206570
min    0.119000   0.000000   0.000000   0.000000   0.000000
25%   0.230000   0.000000   0.000000   0.000000   0.000000
50%   0.402500   0.000000   0.000000   0.000000   0.000000
75%   0.897100   0.000000   0.000000   0.000000   0.000000
max   1.000000   1.000000   1.000000   1.000000   1.000000

      Dirtiness  Bumps  Class
count  1613.000000  1613.000000  1613.000000
```

```

mean      0.034098    0.243025    1.364538
std       0.181537    0.429043    0.481450
min       0.000000    0.000000    1.000000
25%      0.000000    0.000000    1.000000
50%      0.000000    0.000000    1.000000
75%      0.000000    0.000000    2.000000
max       1.000000    1.000000    2.000000

```

[8 rows x 34 columns]

[17]: df\_clean.head()

	X_Minimum	X_Maximum	Y_Minimum	Y_Maximum	Pixels_Areas	X_Perimeter	Y_Perimeter	Sum_of_Luminosity	Minimum_of_Luminosity	Maximum_of_Luminosity	Orientation_Index	Luminosity_Index	SigmoidOfAreas	Pastry	Z_Scratch	K_Scratch	Stains	Dirtiness	Bumps	Class	
0	42	50	270900	270944	267.0	17.0	44.0	24220.0	76	108	...	0.8182	-0.2913	0.5822	1	0	0	0	0	0	1
1	645	651	2538079	2538108	108.0	10.0	30.0	11397.0	84	123	...	0.7931	-0.1756	0.2984	1	0	0	0	0	0	1
2	829	835	1553913	1553931	71.0	8.0	19.0	7972.0	99	125	...	0.6667	-0.1228	0.2150	1	0	0	0	0	0	1
3	853	860	369370	369415	176.0	13.0	45.0	18996.0	99	126	...	0.8444	-0.1568	0.5212	1	0	0	0	0	0	1
4	1289	1306	498078	498335	2409.0	60.0	260.0	246930.0	37	126	...	0.9338	-0.1992	1.0000	1	0	0	0	0	0	1

[5 rows x 34 columns]

[ ]: