

Data Collection and Preprocessing Phase

Date	21 June 2024
Team ID	740005
Project Title	Estimating Presence or Absence of smoking Through Bio Signals
Maximum Marks	6 Marks

Data Exploration and Preprocessing Report

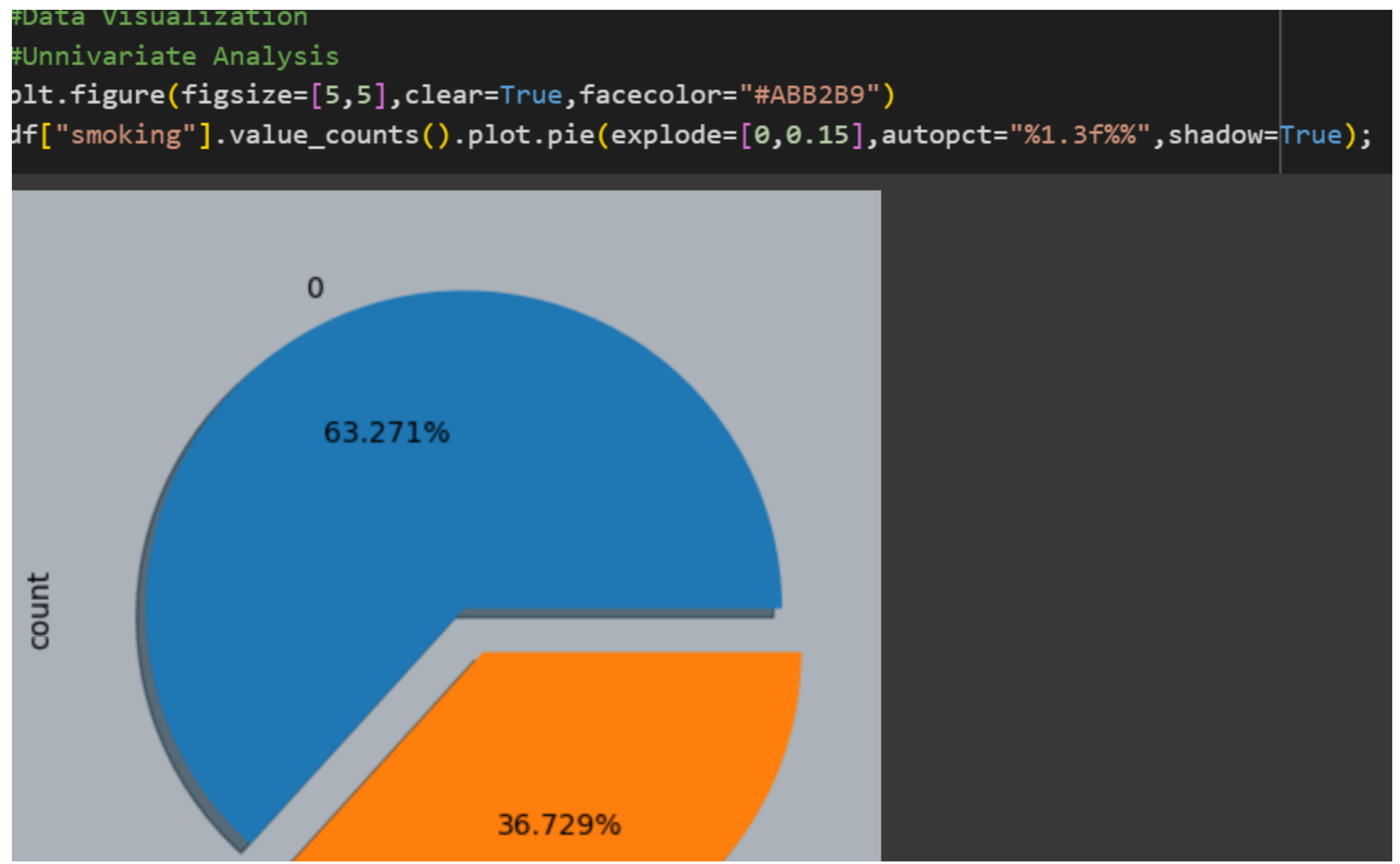
Effective data exploration and preprocessing are foundational steps in developing a reliable system for estimating smoking behavior using biosignals. By understanding the characteristics of the data, addressing noise and outliers, and extracting meaningful features, the processed data is now ready for further analysis and model development.

Section	Description																																																																																																						
Data Overview	<div><div>Descriptive statistics:</div><div><pre>df.describe()</pre><table><tr><th></th><th>ID</th><th>gender</th><th>age</th><th>height(cm)</th><th>weight(kg)</th><th>waist(cm)</th><th>eyesight(left)</th><th>eyesight(right)</th><th>hearing(left)</th><th>hearing(right)</th></tr><tr><td>count</td><td>55692.000000</td><td>55692.000000</td><td>55692.000000</td><td>55692.000000</td><td>55692.000000</td><td>55692.000000</td><td>55692.000000</td><td>55692.000000</td><td>55692.000000</td></tr><tr><td>mean</td><td>27845.500000</td><td>0.635657</td><td>44.182917</td><td>164.649321</td><td>65.864936</td><td>82.046418</td><td>1.012623</td><td>1.007443</td><td>1.025587</td></tr><tr><td>std</td><td>16077.039933</td><td>0.481250</td><td>12.071418</td><td>9.194597</td><td>12.820306</td><td>9.274223</td><td>0.486873</td><td>0.485964</td><td>0.157902</td></tr><tr><td>min</td><td>0.000000</td><td>0.000000</td><td>20.000000</td><td>130.000000</td><td>30.000000</td><td>51.000000</td><td>0.100000</td><td>0.100000</td><td>1.000000</td></tr><tr><td>25%</td><td>13922.750000</td><td>0.000000</td><td>40.000000</td><td>160.000000</td><td>55.000000</td><td>76.000000</td><td>0.800000</td><td>0.800000</td><td>1.000000</td></tr><tr><td>50%</td><td>27845.500000</td><td>1.000000</td><td>40.000000</td><td>165.000000</td><td>65.000000</td><td>82.000000</td><td>1.000000</td><td>1.000000</td><td>1.000000</td></tr><tr><td>75%</td><td>41768.250000</td><td>1.000000</td><td>55.000000</td><td>170.000000</td><td>75.000000</td><td>88.000000</td><td>1.200000</td><td>1.200000</td><td>1.000000</td></tr><tr><td>max</td><td>55691.000000</td><td>1.000000</td><td>85.000000</td><td>190.000000</td><td>135.000000</td><td>129.000000</td><td>9.900000</td><td>9.900000</td><td>2.000000</td></tr><tr><td colspan="11">8 rows x 27 columns</td></tr></table></div></div>		ID	gender	age	height(cm)	weight(kg)	waist(cm)	eyesight(left)	eyesight(right)	hearing(left)	hearing(right)	count	55692.000000	55692.000000	55692.000000	55692.000000	55692.000000	55692.000000	55692.000000	55692.000000	55692.000000	mean	27845.500000	0.635657	44.182917	164.649321	65.864936	82.046418	1.012623	1.007443	1.025587	std	16077.039933	0.481250	12.071418	9.194597	12.820306	9.274223	0.486873	0.485964	0.157902	min	0.000000	0.000000	20.000000	130.000000	30.000000	51.000000	0.100000	0.100000	1.000000	25%	13922.750000	0.000000	40.000000	160.000000	55.000000	76.000000	0.800000	0.800000	1.000000	50%	27845.500000	1.000000	40.000000	165.000000	65.000000	82.000000	1.000000	1.000000	1.000000	75%	41768.250000	1.000000	55.000000	170.000000	75.000000	88.000000	1.200000	1.200000	1.000000	max	55691.000000	1.000000	85.000000	190.000000	135.000000	129.000000	9.900000	9.900000	2.000000	8 rows x 27 columns										
		ID	gender	age	height(cm)	weight(kg)	waist(cm)	eyesight(left)	eyesight(right)	hearing(left)	hearing(right)																																																																																												
	count	55692.000000	55692.000000	55692.000000	55692.000000	55692.000000	55692.000000	55692.000000	55692.000000	55692.000000																																																																																													
	mean	27845.500000	0.635657	44.182917	164.649321	65.864936	82.046418	1.012623	1.007443	1.025587																																																																																													
	std	16077.039933	0.481250	12.071418	9.194597	12.820306	9.274223	0.486873	0.485964	0.157902																																																																																													
	min	0.000000	0.000000	20.000000	130.000000	30.000000	51.000000	0.100000	0.100000	1.000000																																																																																													
	25%	13922.750000	0.000000	40.000000	160.000000	55.000000	76.000000	0.800000	0.800000	1.000000																																																																																													
	50%	27845.500000	1.000000	40.000000	165.000000	65.000000	82.000000	1.000000	1.000000	1.000000																																																																																													
	75%	41768.250000	1.000000	55.000000	170.000000	75.000000	88.000000	1.200000	1.200000	1.000000																																																																																													
	max	55691.000000	1.000000	85.000000	190.000000	135.000000	129.000000	9.900000	9.900000	2.000000																																																																																													
8 rows x 27 columns																																																																																																							

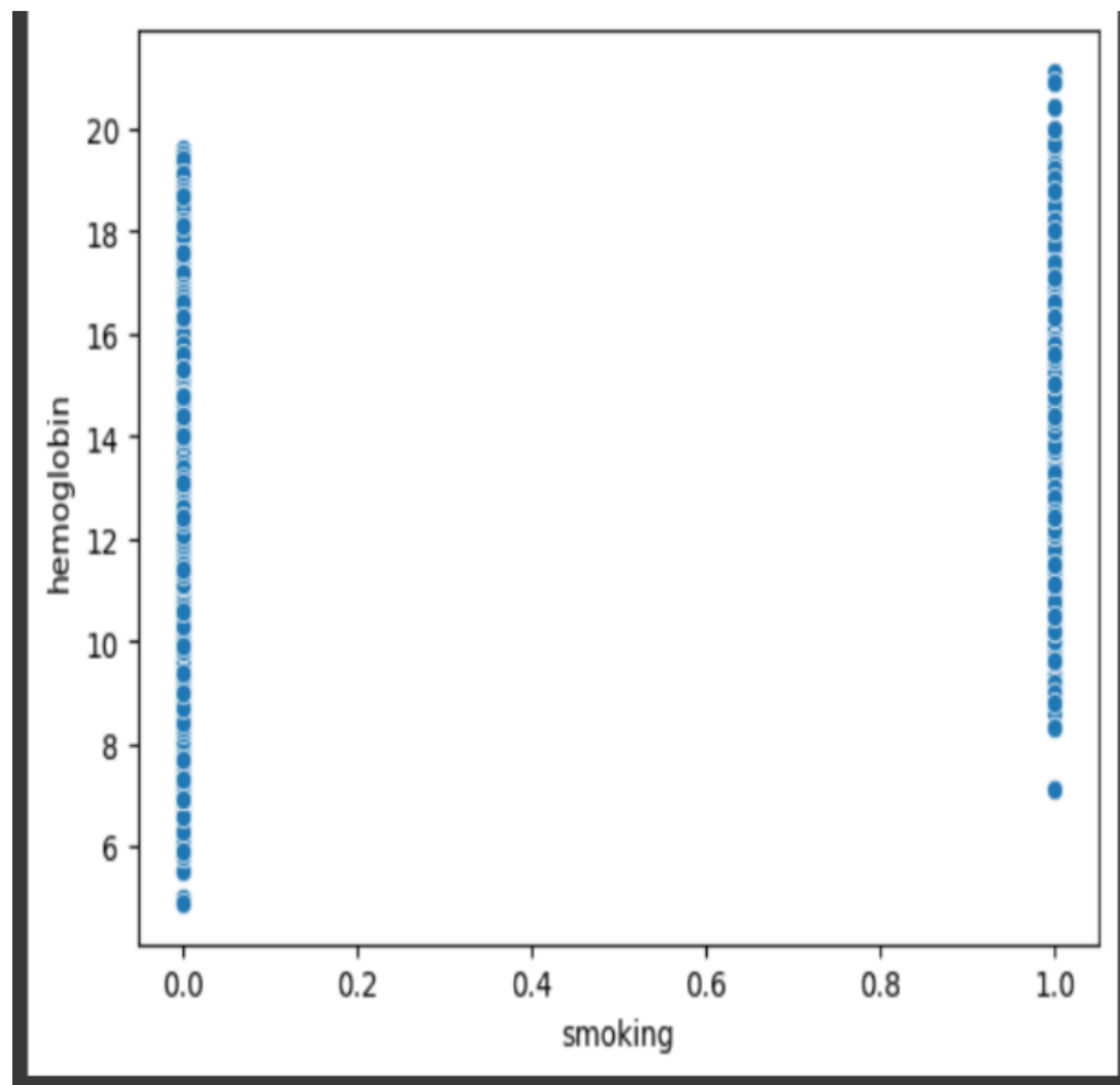
hearing(left)	hearing(right)	...	hemoglobin	Urine_protein	serum_creatinine	AST	ALT	Gtp	oral	dental_caries
55692.000000	55692.000000	...	55692.000000	55692.000000	55692.000000	55692.000000	55692.000000	55692.000000	55692.0	55692.000000
1.025587	1.026144	...	14.622592	1.087212	0.885738	26.182935	27.036037	39.952201	0.0	0.213334
0.157902	0.159564	...	1.564498	0.404882	0.221524	19.355460	30.947853	50.290539	0.0	0.409665
1.000000	1.000000	...	4.900000	1.000000	0.100000	6.000000	1.000000	1.000000	0.0	0.000000
1.000000	1.000000	...	13.600000	1.000000	0.800000	19.000000	15.000000	17.000000	0.0	0.000000
1.000000	1.000000	...	14.800000	1.000000	0.900000	23.000000	21.000000	25.000000	0.0	0.000000
1.000000	1.000000	...	15.800000	1.000000	1.000000	28.000000	31.000000	43.000000	0.0	0.000000
2.000000	2.000000	...	21.100000	6.000000	11.600000	1311.000000	2914.000000	999.000000	0.0	1.000000

dental_caries	tartar	smoking
55692.000000	55692.000000	55692.000000
0.213334	0.555556	0.367288
0.409665	0.496908	0.482070
0.000000	0.000000	0.000000
0.000000	0.000000	0.000000
0.000000	1.000000	0.000000
0.000000	1.000000	1.000000
1.000000	1.000000	1.000000

Univariate
Analysis



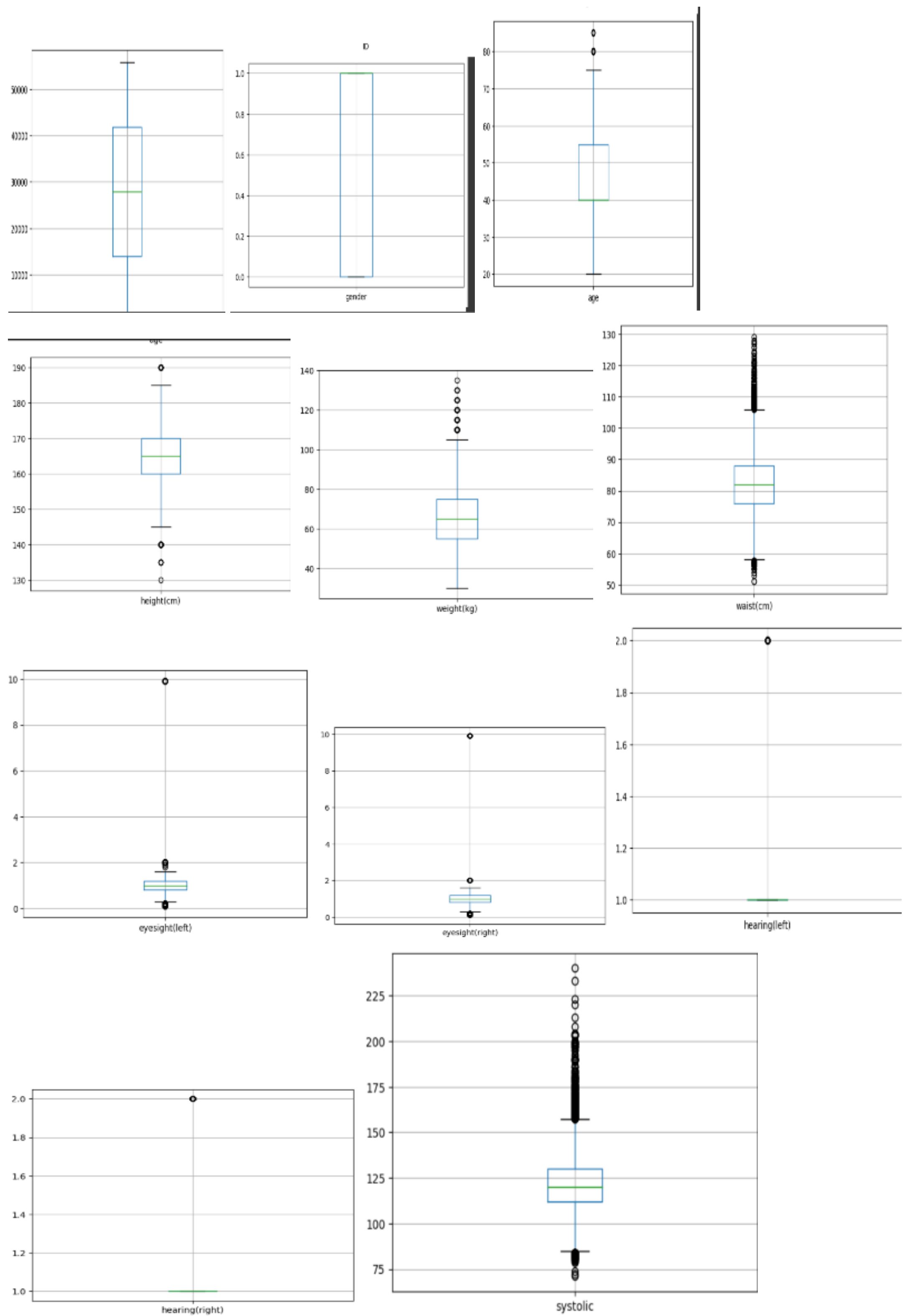
Bivariate Analysis

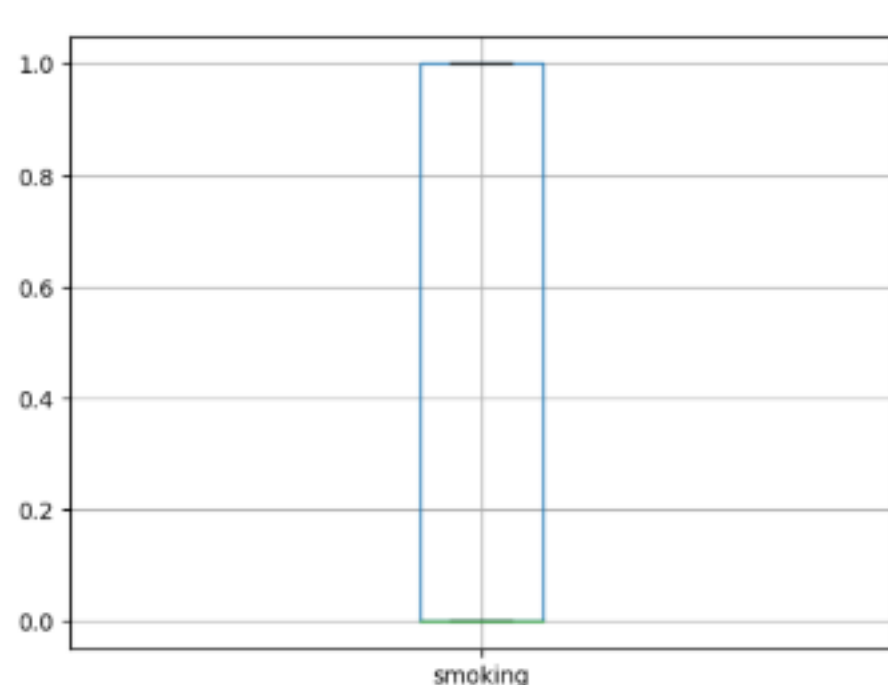
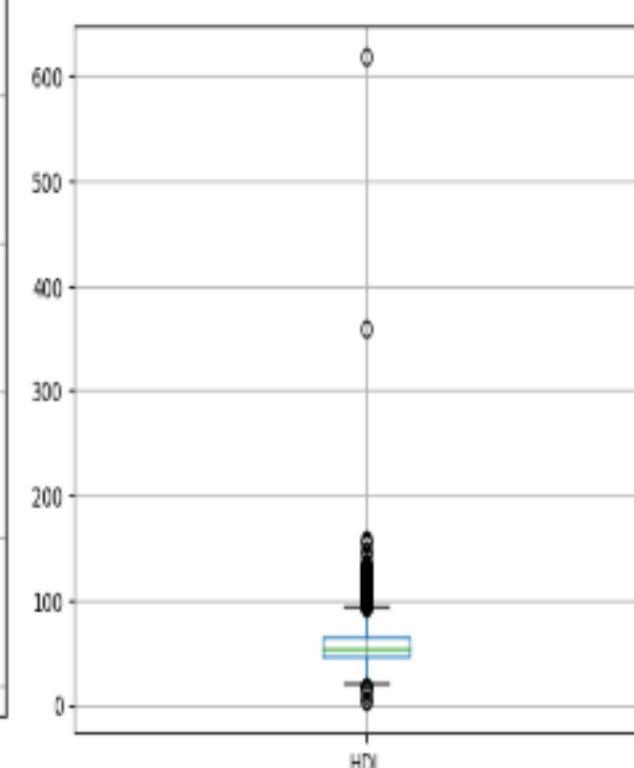
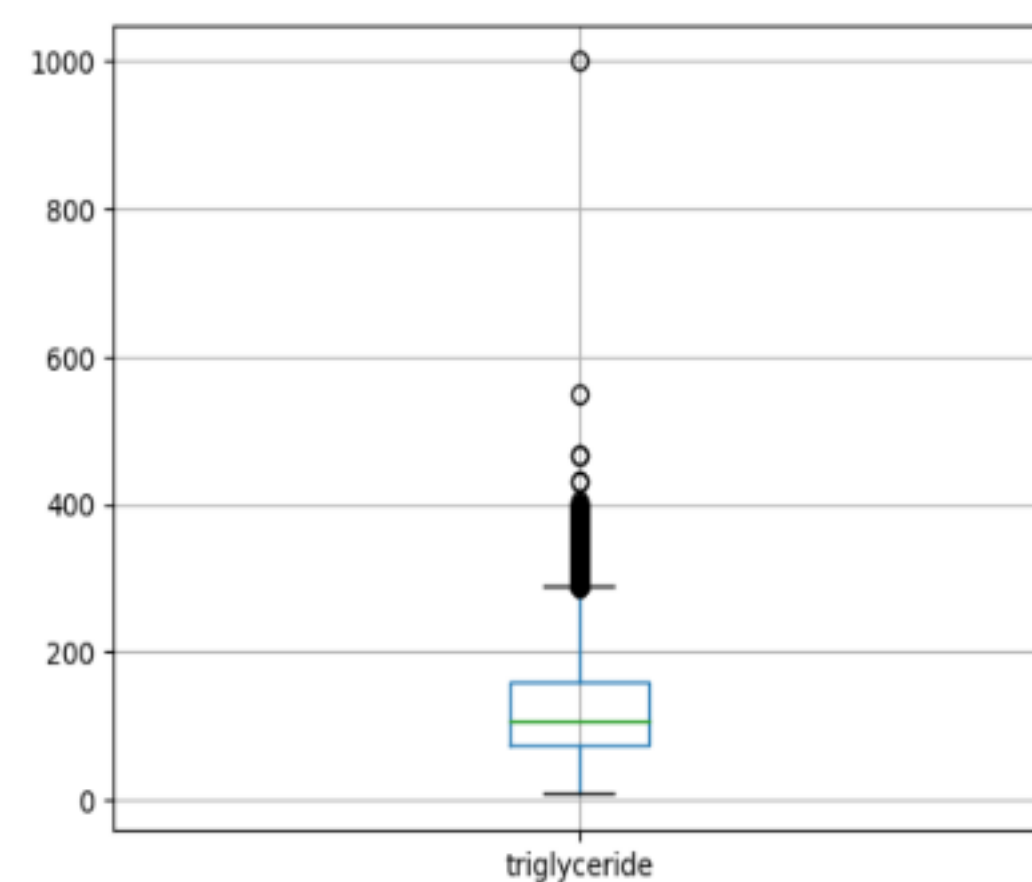
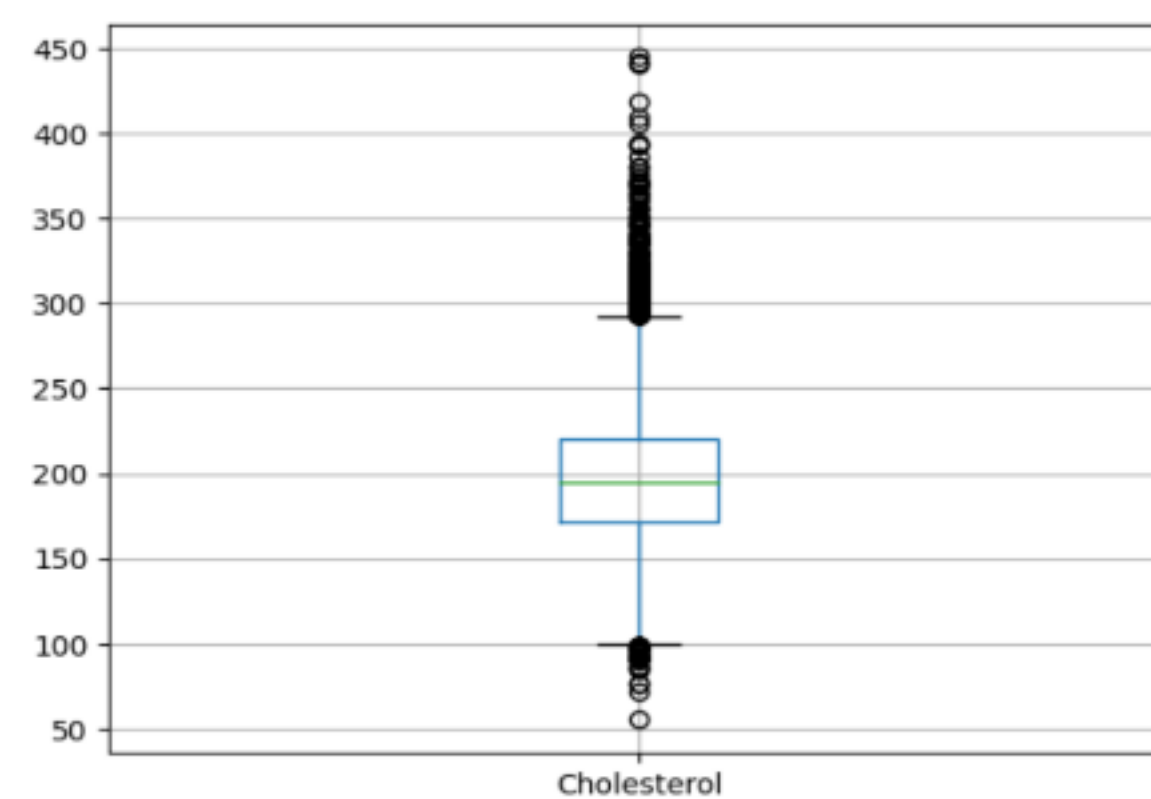
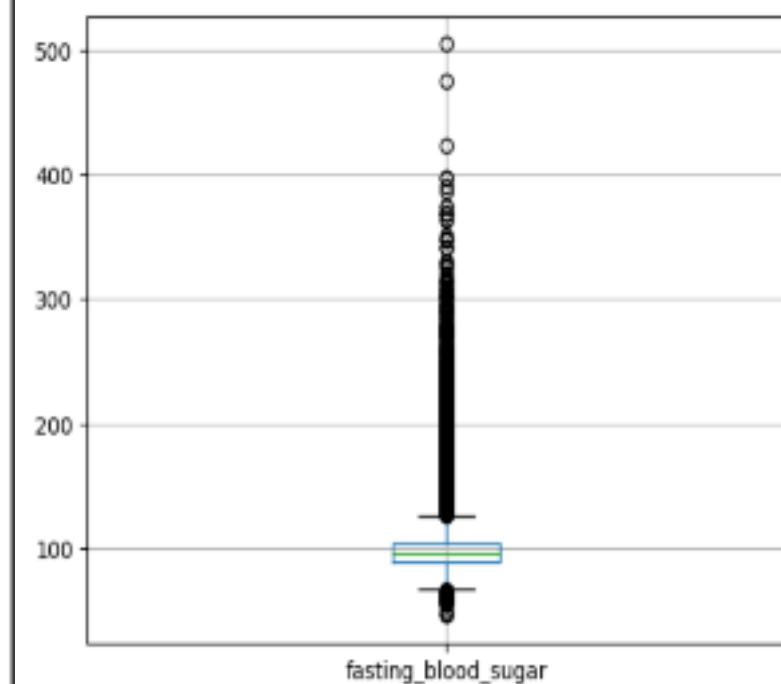
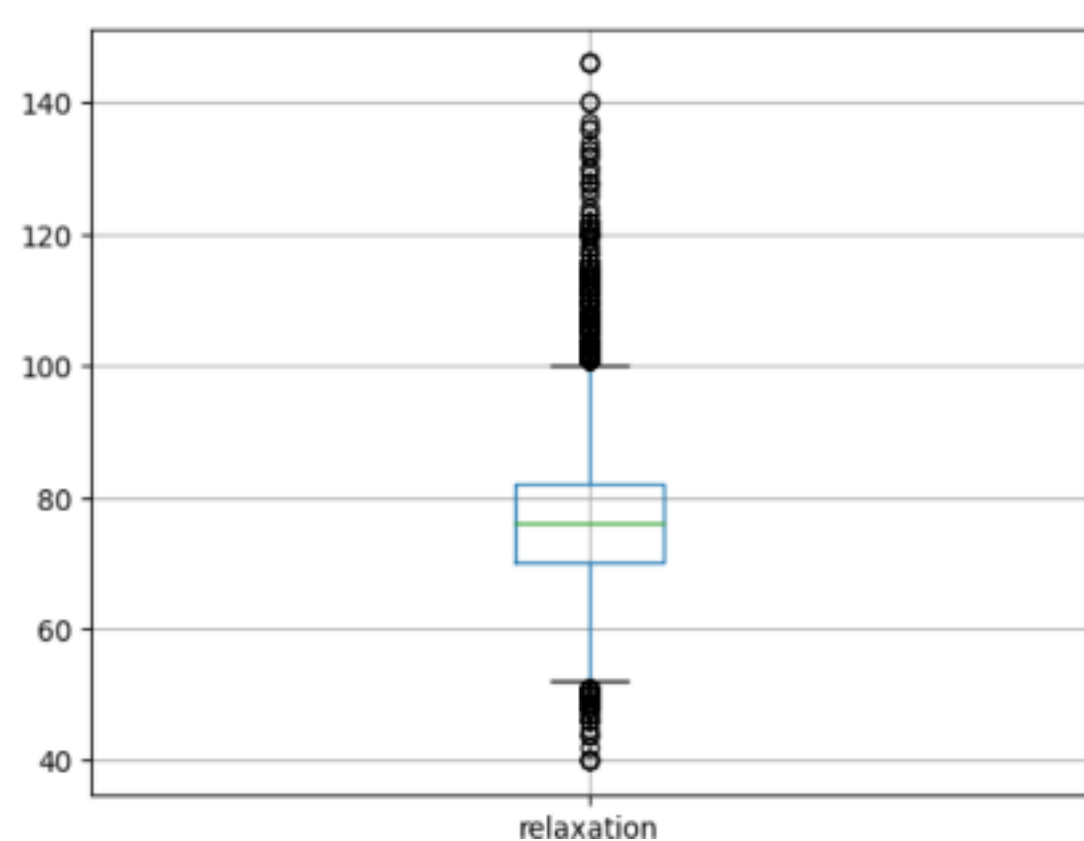


Multivariate Analysis

-

Outliers and Anomalies





Data Preprocessing Code Screenshots

Loading Data

```
ID gender age height(cm) weight(kg) waist(cm) eyesight(left) eyesight(right) hearing(left) hearing(right) ... hemoglobin p
0 0 F 40 155 60 81.3 1.2 1.0 1.0 1.0 . 12.9
1 1 F 40 160 60 81.0 0.8 0.6 1.0 1.0 . 12.7
2 2 M 55 170 60 80.0 0.8 0.8 1.0 1.0 . 15.3
3 3 M 40 165 70 88.0 1.5 1.5 1.0 1.0 . 14.7
4 4 F 40 155 60 86.0 1.0 1.0 1.0 1.0 . 12.5
... ..
55687 55675 F 40 170 65 75.0 0.9 0.9 1.0 1.0 . 12.3
55688 55681 F 45 160 50 70.0 1.2 1.2 1.0 1.0 . 14.0
55689 55683 F 55 160 50 68.5 1.0 1.2 1.0 1.0 . 12.4
55690 55684 M 50 165 60 78.0 0.8 1.0 1.0 1.0 . 14.4
55691 55691 M 55 160 65 85.0 0.9 0.7 1.0 1.0 . 15.0
55692 rows x 27 columns
```

hemoglobin	Urine protein	serum creatinine	AST	ALT	Gtp	oral	dental caries	tartar	smoking
12.9	1.0	0.7	18.0	19.0	27.0	Y	0	Y	0
12.7	1.0	0.6	22.0	19.0	18.0	Y	0	Y	0
15.8	1.0	1.0	21.0	16.0	22.0	Y	0	N	1
14.7	1.0	1.0	19.0	26.0	18.0	Y	0	Y	0
12.5	1.0	0.6	16.0	14.0	22.0	Y	0	N	0
...
12.3	1.0	0.6	14.0	7.0	10.0	Y	1	Y	0
14.0	1.0	0.9	20.0	12.0	14.0	Y	0	Y	0
12.4	1.0	0.5	17.0	11.0	12.0	Y	0	N	0
14.4	1.0	0.7	20.0	19.0	18.0	Y	0	N	0
15.0	1.0	0.8	26.0	29.0	41.0	Y	0	Y	1

Handling Missing Data

```
ID 0
gender 0
age 0
height(cm) 0
weight(kg) 0
waist(cm) 0
eyesight(left) 0
eyesight(right) 0
hearing(left) 0
hearing(right) 0
systolic 0
relaxation 0
fasting_blood_sugar 0
Cholesterol 0
triglyceride 0
HDL 0
LDL 0
hemoglobin 0
Urine_protein 0
serum_creatinine 0
AST 0
ALT 0
Gtp 0
oral 0
dental_caries 0
tartar 0
smoking 0
dtype: int64
```

Data Transformati on

-

Feature Engineering	Attached the codes in final submission.
Save Processed Data	-