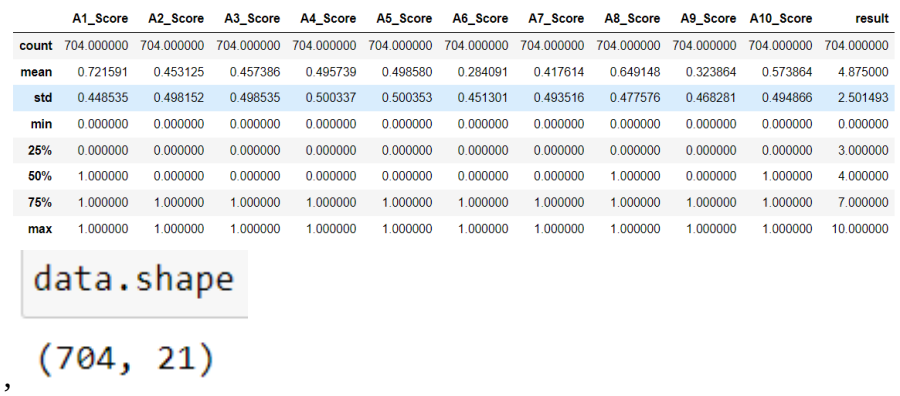
Data Collection and Preprocessing Phase

Date	15 March 2024
Team ID	SWTID1720000747
Project Title	Detection Of Autistic Spectrum Disorder: Classification
Maximum Marks	6 Marks

Data Exploration and Preprocessing Template

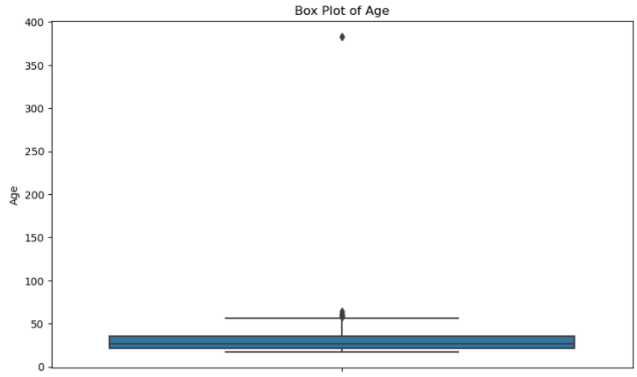
Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

Section	Description
Data Overview	
Univariate Analysis	Box-plot – age

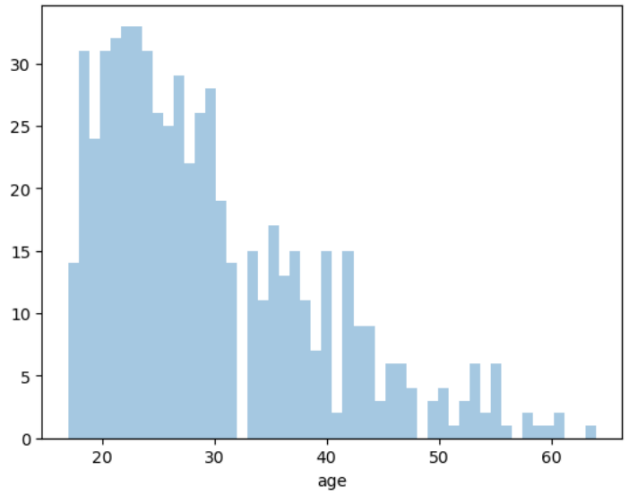
	A1_Score	A2_Score	A3_Score	A4_Score	A5_Score	A6_Score	A7_Score	A8_Score	A9_Score	A10_Score	result
count	704.000000	704.000000	704.000000	704.000000	704.000000	704.000000	704.000000	704.000000	704.000000	704.000000	704.000000
mean	0.721591	0.453125	0.457386	0.495739	0.498580	0.284091	0.417614	0.649148	0.323864	0.573864	4.875000
std	0.448535	0.498152	0.498535	0.500337	0.500353	0.451301	0.493516	0.477576	0.468281	0.494866	2.501493
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	3.000000
50%	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000	1.000000	4.000000
75%	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	7.000000
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	10.000000

```
data.shape
```

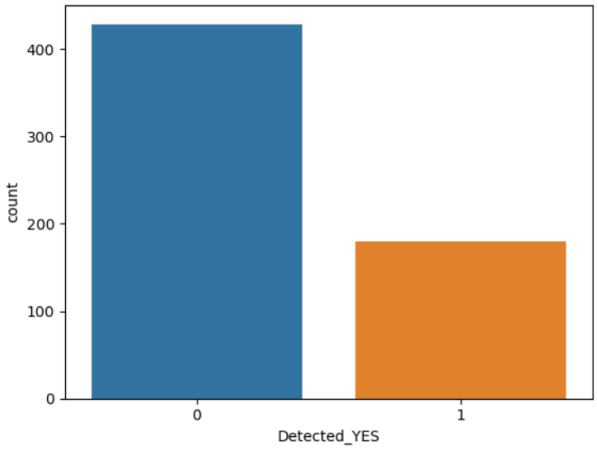
```
(704, 21)
```

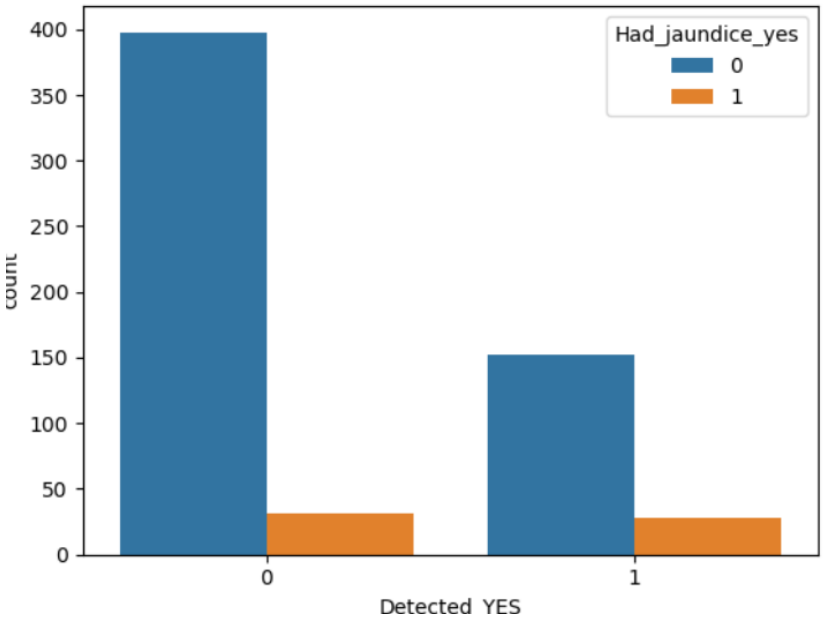
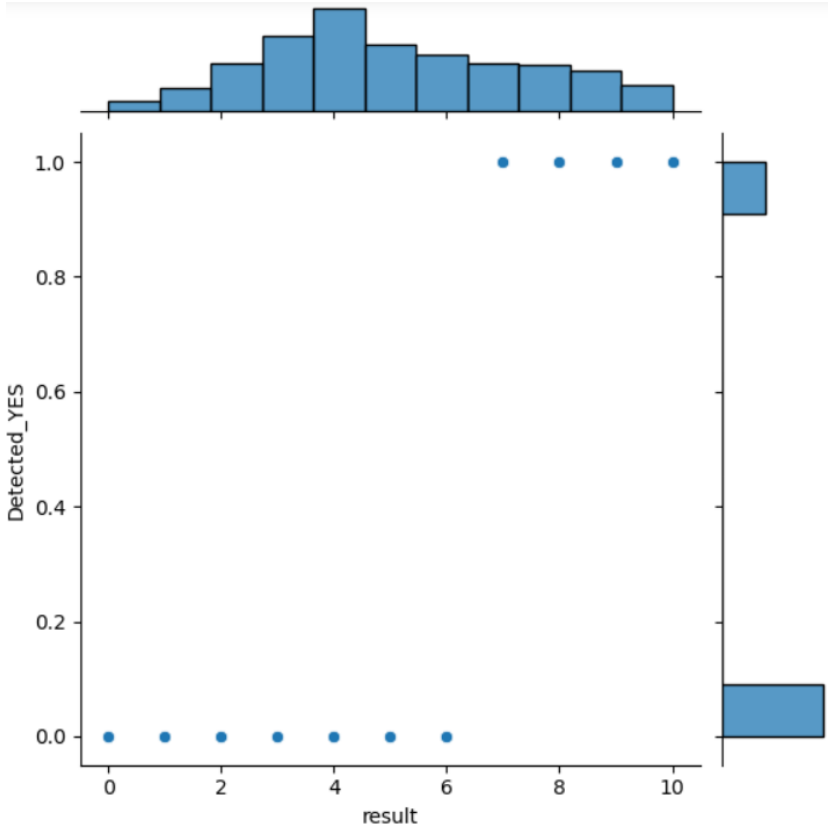


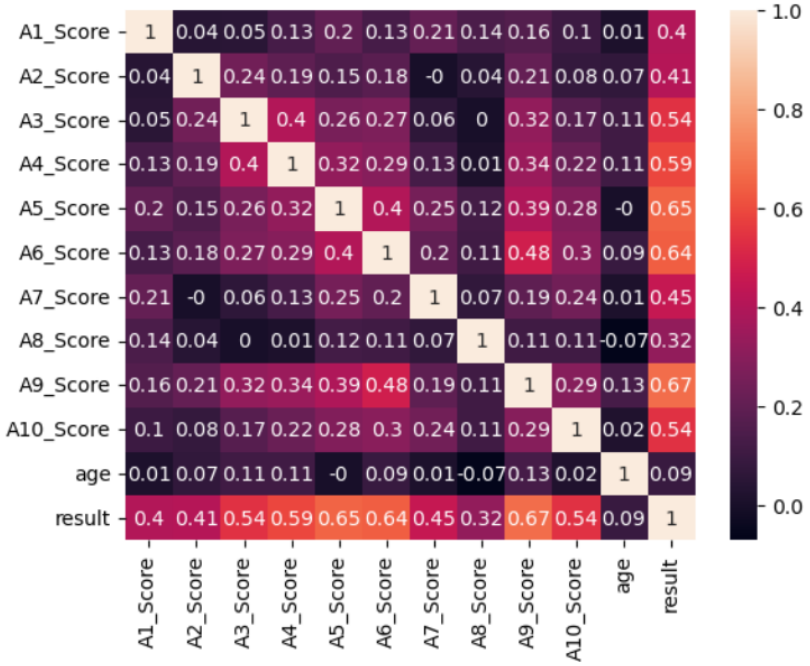
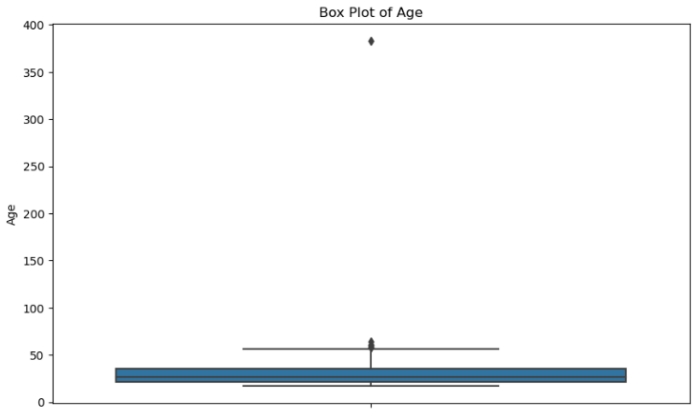
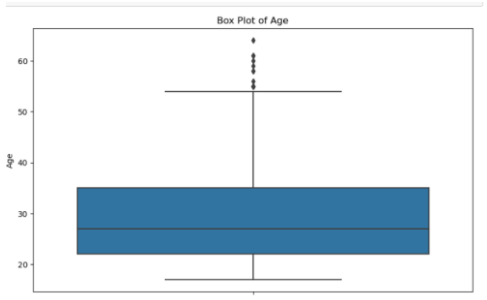
Distplot – age (after removing outliers)



Bar plot – Detected_YES



Bivariate Analysis	Countplot – Had_jaundice_YES and Detected_YES  Jointplot – result and Detected_YES 

Multivariate Analysis	Heat map 
Outliers and Anomalies	Box-plot – age  

Data Preprocessing Code Screenshots

Loading Data

```
data = pd.read_csv("Autism_Data.csv")
```

```
data
```

	A1_Score	A2_Score	A3_Score	A4_Score	A5_Score	A6_Score	A7_Score	A8_Score	A9_Score	A10_Score	...	gender
0	1	1	1	1	0	0	1	1	0	0	...	f
1	1	1	0	1	0	0	0	1	0	1	...	m
2	1	1	0	1	1	0	1	1	1	1	...	m
3	1	1	0	1	0	0	1	1	0	1	...	f
4	1	0	0	0	0	0	0	1	0	0	...	f

Handling Missing Data

```
data.isnull().sum()
```

```
data.replace('?',np.nan,inplace=True)
```

```
(data['age'].eq('?')).any()
```

```
data_p = data
data_p.dropna(inplace=True)
```

```
data_t = data_p[data_p['age']!=383]
data_t['age'].mean()
```

```
29.63486842105263
```

```
data.loc[data.age == 383, 'age'] = 30
data['age'] = data['age'].fillna(30)
```

Data Transformation

```
from sklearn.preprocessing import StandardScaler
sc=StandardScaler()
x_s=sc.fit_transform(X)
X = pd.DataFrame(x_s)
```

Feature Engineering	`sex = pd.get_dummies(data['gender'],drop_first=True)`
	`jaund = pd.get_dummies(data['jundice'],drop_first=True,prefix="Had_jaundice")`
	`rel_autism = pd.get_dummies(data['austim'],drop_first=True, prefix = "Rel_had")`
	`detected = pd.get_dummies(data['Class/ASD'],drop_first=True,prefix="Detected")`

Save Processed Data	`data_featured = pd.concat([data,sex,jaund,rel_autism,detected],axis=1)` `data_featured.head()`