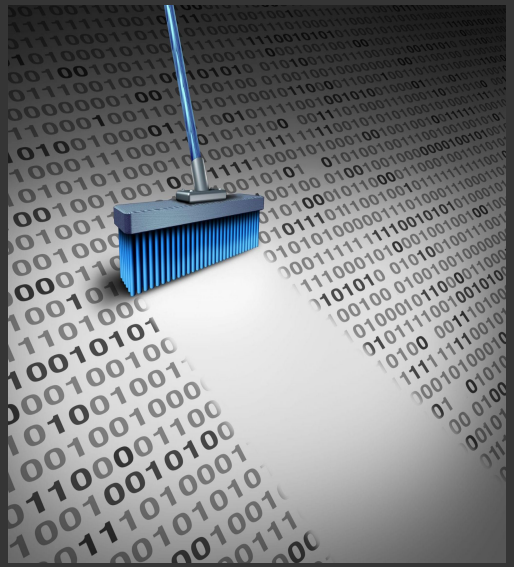
01 What the data is about

ANALYST MEETING
ANALYST MEETING
ANALYST MEETING

Table of contents



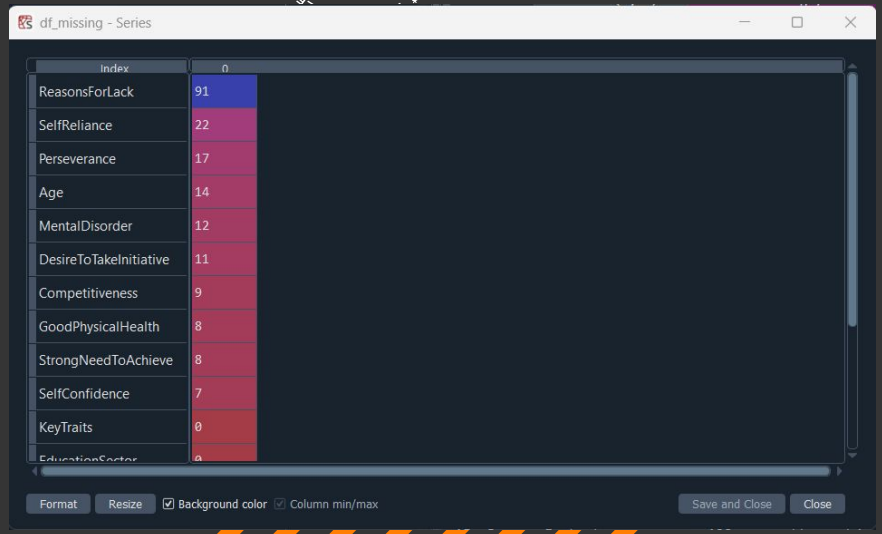
“Entrepreneurial competency in university students”

	EducationSector	Target IndividualProject	Age	Gender	City	Influenced	Perseverance	DesireToTakeInitiative
190	Economic Sciences, Business Studies, Commerce and Law	Yes	20	Female	Yes	No	4	
191	Economic Sciences, Business Studies, Commerce and Law	No	18	Female	Yes	No	4	
192	Engineering Sciences	No	17	Male	No	Yes	3	
193	Others	Yes	21	Female	Yes	No	4	
194	Others	Yes	19	Female	No	Yes	4	
195	Art, Music or Design	Yes	20	Male	Yes	Yes	2	
196	Engineering Sciences	No	20	Male	Yes	Yes	4	
197	Engineering Sciences	No	19	Male	Yes	yes	5	
198	Engineering Sciences	Yes	19	Female	Yes	No	3	
199	Engineering Sciences	Yes	20	Male	Yes	Yes	3	
200	Art, Music or Design	Yes	19	Male	Yes	Yes		
201	Engineering Sciences	No	19	Male	Yes	No		
202	Engineering Sciences	Yes	20	Male	Yes	yes		
203	Engineering Sciences	No	19	Male	no	Yes		
204	Mathematics or Natural Sciences	Yes	17	Male	Yes	Yes		
205	Art, Music or Design	Yes	20	Female	Yes	Yes		
206	Engineering Sciences	Yes	20	Female	Yes	unknown		
207	Engineering Sciences	Yes	20	Male	Yes	Yes	4	
208	Engineering Sciences	Yes	21	Male	Yes	Yes		
209	Engineering Sciences	Yes	21	Female	No	No		
210	Engineering Sciences	No	20	Male	Yes	No		
211	Economic Sciences, Business Studies, Commerce and Law	Yes	19	Male	Yes	No		
212	Engineering Sciences	Yes	20	Female	Yes	Yes		
213	Economic Sciences, Business Studies, Commerce and Law	No	19	Male	Yes	Yes		

Page 5 Page 10

“

Entrepreneurial competency in university students”



Database informations

- 17 columns for 219 rows

- 9 numerical columns

- 8 non numerical values

2022

02 DATA CLEANING

MEETING

missing values

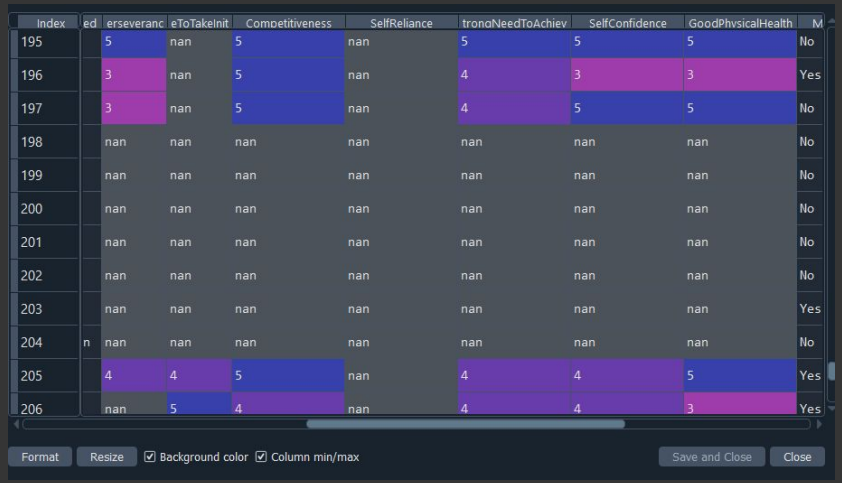
Calculate the number of missing values.

```
df_missing = data.isna().sum().sort_values(ascending=False)
df_missing
```

✓ 0.7s

```
ReasonsForLack              91
SelfReliance                22
Perseverance                17
Age                         14
MentalDisorder              12
DesireToTakeInitiative      11
Competitiveness              9
GoodPhysicalHealth           8
StrongNeedToAchieve          8
SelfConfidence               7
KeyTraits                    0
EducationSector              0
Target IndividualProject     0
Influenced                   0
City                         0
Gender                       0
Target-ent_competency        0
dtype: int64
```

missing values



Index	ed	erseveranc	eToTakeInit	Competitiveness	SelfReliance	tronaNeedToAchiev	SelfConfidence	GoodPhysicalHealth	M
195		5	nan	5	nan	5	5	5	No
196		3	nan	5	nan	4	3	3	Yes
197		3	nan	5	nan	4	5	5	No
198		nan	nan	nan	nan	nan	nan	nan	No
199		nan	nan	nan	nan	nan	nan	nan	No
200		nan	nan	nan	nan	nan	nan	nan	No
201		nan	nan	nan	nan	nan	nan	nan	No
202		nan	nan	nan	nan	nan	nan	nan	No
203		nan	nan	nan	nan	nan	nan	nan	Yes
204	n	nan	nan	nan	nan	nan	nan	nan	No
205		4	4	5	nan	4	4	5	Yes
206		nan	5	4	nan	4	4	3	Yes

Format | Resize | ☑ Background color | ☑ Column min/max | Save and Close | Close

Drop all the rows with more than 3 missing values

```
data= data[data.isnull().sum(axis=1) < 3]
```

✓ 0.5s

missing values

```
modepers = data['Perseverance'].mode()
data['Perseverance'] = np.where(data['Perseverance'].isnull(), modepers , data['Perseverance'])
```

✓ 0.8s

check inconsistency

display the unique values of each column

```
for i in data.columns:
    print(data[i].unique())
```

✓ 0.6s

```
Output exceeds the size limit. Open the full output data in a text editor
['Engineering Sciences' 'Others'
 'Economic Sciences, Business Studies, Commerce and Law'
 'Art, Music or Design' 'Humanities and Social Sciences'
 'Medicine, Health Sciences' 'Teaching Degree (e.g., B.Ed)'
 'Mathematics or Natural Sciences' 'Language and Cultural Studies']
['No' 'Yes']
[19. 22. 18. 20. 17. 21. 23. 26. 24. nan 25.]
['Male' 'Female' 'male']
['Yes' 'No' 'unknown' 'no']
['No' 'Yes' 'yes' 'unknown']
[ 2.  3.  4.  1.  5. nan]
[ 2.  3.  4.  1.  5. nan]
[ 3.  4.  2.  5.  1. nan]
[ 3.  4.  1.  2.  5. nan]
[ 2.  4.  3.  5.  1. nan]
[ 2.  3.  4.  1.  5. nan]
[ 3.  4.  2.  1.  5. nan]
['Yes' 'No' nan]
['Passion' 'Vision' 'Rrresilience' 'Positivity' 'Work Ethic' 'Resilience'
 'passion']
```

check inconsistency

```
#there is "male" in lowercase in the column "gender"
data['Gender'] = data['Gender'].str.capitalize()
```

✓ 0.3s

```
C:\Users\kyrie\AppData\Local\Temp\ipykernel_14908\2188931611.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/ir
  data['Gender'] = data['Gender'].str.capitalize()
```

```
#data['KeyTraits'] = data['KeyTraits'].str.capitalize()
data['KeyTraits'] = np.where(data['KeyTraits'] == "passion", "Passion" , data['KeyTraits'])
```

✓ 0.4s

check Duplicates

removing duplicates

```
print(data.duplicated())
```

✓ 0.4s

```
0      False
1      False
2      False
3      False
4      False
       ...
214    False
215    False
216    False
217    False
218    False
Length: 203, dtype: bool
```

```
data.drop_duplicates(inplace = True)
```

✓ 0.4s

VSCODE recap step by step

2022

03 DATA ANALYSIS

MEETING

Improvement axe

Creation of a composite indicator with all the column to define what are the most ‘entrepreneurial ready’ profiles


Q&A


Thanks!