Automated EDA on Visa dataset

Linkedin: https://www.linkedin.com/in/satya-nerurkar-9b0655190/ (https://www.linkedin.com/in/satya-nerurkar-9b0655190/)
Github: https://github.com/SatyaNerurkar (https://github.com/SatyaNerurkar)

In [1]:
```
# Install supporting libaries
!pip install pandas-profiling sweetviz --upgrade --quiet
```

In [2]:
```
# Import supporting libaries
import pandas as pd
from pandas_profiling import ProfileReport
import sweetviz as sv
import dtale
```

EDA using pandas profiling.

In [3]:
```
# EDA using pandas-profiling
profile = ProfileReport(pd.read_csv(".\Visadataset.csv"),
                        explorative=True)
```

In [4]:
```
# Saving results to a HTML file
profile.to_file("pandas_profiling_visa_report.html")

# Saving results to a JSON file
profile.to_file("pandas_profiling_visa_report.json")
```

Summarize dataset: 100% 35/35 [00:17<00:00, 2.01it/s, Completed]

Generate report structure: 100% 1/1 [00:08<00:00, 8.33s/it]

Render HTML: 100% 1/1 [00:02<00:00, 2.43s/it]

Export report to file: 100% 1/1 [00:00<00:00, 5.80it/s]

Render JSON: 100% 1/1 [00:01<00:00, 1.59s/it]

Export report to file: 100% 1/1 [00:00<00:00, 10.50it/s]

In [5]:
```
# Displaying the report as a set of widgets
profile.to_widgets()
```

Overview | Variables | Interactions | Correlations | Missing value: | Sample

Overview | Alerts (3) | Reproduction

Number of variables	12	Categorical	6
		Boolean	3
Number of observations	25480	Numeric	3
Missing cells	0		
Missing cells (%)	0.0%		
Duplicate rows	0		
Duplicate rows (%)	0.0%		
Total size in memory	14.1 MiB		
Average record size in memory	581.9 B		

Report generated by YData (https://ydata.ai/?utm_source=opensource&utm_medium=pandasprofiling&utm_campaign=report).

In [6]:
```
# Displaying report in notebook cell.
profile.to_notebook_iframe()
```

Overview

Dataset statistics

Number of variables	12
Number of observations	25480
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	14.1 MiB
Average record size in memory	581.9 B

Variable types

Categorical	6
Boolean	3
Numeric	3

Alerts

`case_id` has a high cardinality: 25480 distinct values	**High cardinality**
`case_id` is uniformly distributed	**Uniform**

EDA using Sweetviz

In [7]:
```
#EDA using Sweetviz
sweet_report = sv.analyze(pd.read_csv(".\Visadataset.csv"))
```

Done! Use 'show' commands to display/save. [100%] 00:01 -> (00:00 left)

In [8]:
```
#Saving results to HTML file
sweet_report.show_html('sweet_visa_report.html')
```

```
Report sweet_visa_report.html was generated! NOTEBOOK/COLAB USERS: the web brow
ser MAY not pop up, regardless, the report IS saved in your notebook/colab file
s.
```

EDA using D-tale

In [9]:
```
# Assigning dataset to variable
df = pd.read_csv('.\Visadataset.csv')
dtale.show(df)
```

▶ 12 / 25480

	case_id	continent	education_of_employee	has_job_experience
0	EZYV01	Asia	High School	N
	EZYV02	Asia	Master's	Y
2	EZYV03	Asia	Bachelor's	N
3	EZYV04	Asia	Bachelor's	N
4	EZYV05	Africa	Master's	Y
5	EZYV06	Asia	Master's	Y
6	EZYV07	Asia	Bachelor's	N
7	EZYV08	North America	Bachelor's	Y
8	EZYV09	Asia	Bachelor's	N
9	EZYV10	Europe	Doctorate	Y
10	EZYV11	Asia	Master's	N
1	EZYV12	Asia	High School	Y
12	EZYV13	Asia	Bachelor's	Y
13	EZYV14	Asia	Bachelor's	Y
14	EZYV15	Asia	Master's	Y
15	EZYV16	Asia	High School	Y
16	EZYV17	Europe	Master's	Y

Out[9]: