EDA Implementation using Automated EDA Tools

Submitted by : Nilutpal Das

In [13]:
```
# Importing CSV dataset with a help of Pandas Library
import pandas as pd
df = pd.read_csv('D:\Data Science\Dataset\data5\Visadataset.csv')
```

In [14]:
```
# head() used to provide top 5 data from dataset.
df.head()
```

Out[14]:

	case_id	continent	education_of_employee	has_job_experience	requires_job_training	no_of_employees	yr_of_estab	region_of_employment
0	EZYV01	Asia	High School	N	N	14513	2007	West
1	EZYV02	Asia	Master's	Y	N	2412	2002	Northeast
2	EZYV03	Asia	Bachelor's	N	Y	44444	2008	West
3	EZYV04	Asia	Bachelor's	N	N	98	1897	West
4	EZYV05	Africa	Master's	Y	N	1082	2005	South

1. Pandas Profiling Tools

Pandas profiling is an open source Python module with which we can quickly do an exploratory data analysis with just a few lines of code.

The pandas_profiling library generates a report having:

- An overview of the dataset.
- Variable properties.
- Interaction of variables.
- Correaltion of variables.
- Sample data.
- Missing values.

In [15]:
```
from pandas_profiling import ProfileReport
profile = ProfileReport(df,explorative=True, title='Report')
profile
```

```
Summarize dataset:   0%|          | 0/5 [00:00<?, ?it/s]
Generate report structure:   0%|          | 0/1 [00:00<?, ?it/s]
Render HTML:   0%|          | 0/1 [00:00<?, ?it/s]
```

Report   Overview   Variables   Interactions   Correlations   Missing values   Sample

Overview

Overview   Alerts 3   Reproduction

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

Variables

case_id
Categorical

Distinct	25480
Distinct (%)	100.0%

EZYV01 1
EZYV16995 1
EZYV16993 1

Out[15]:

2. SweetViz Tools

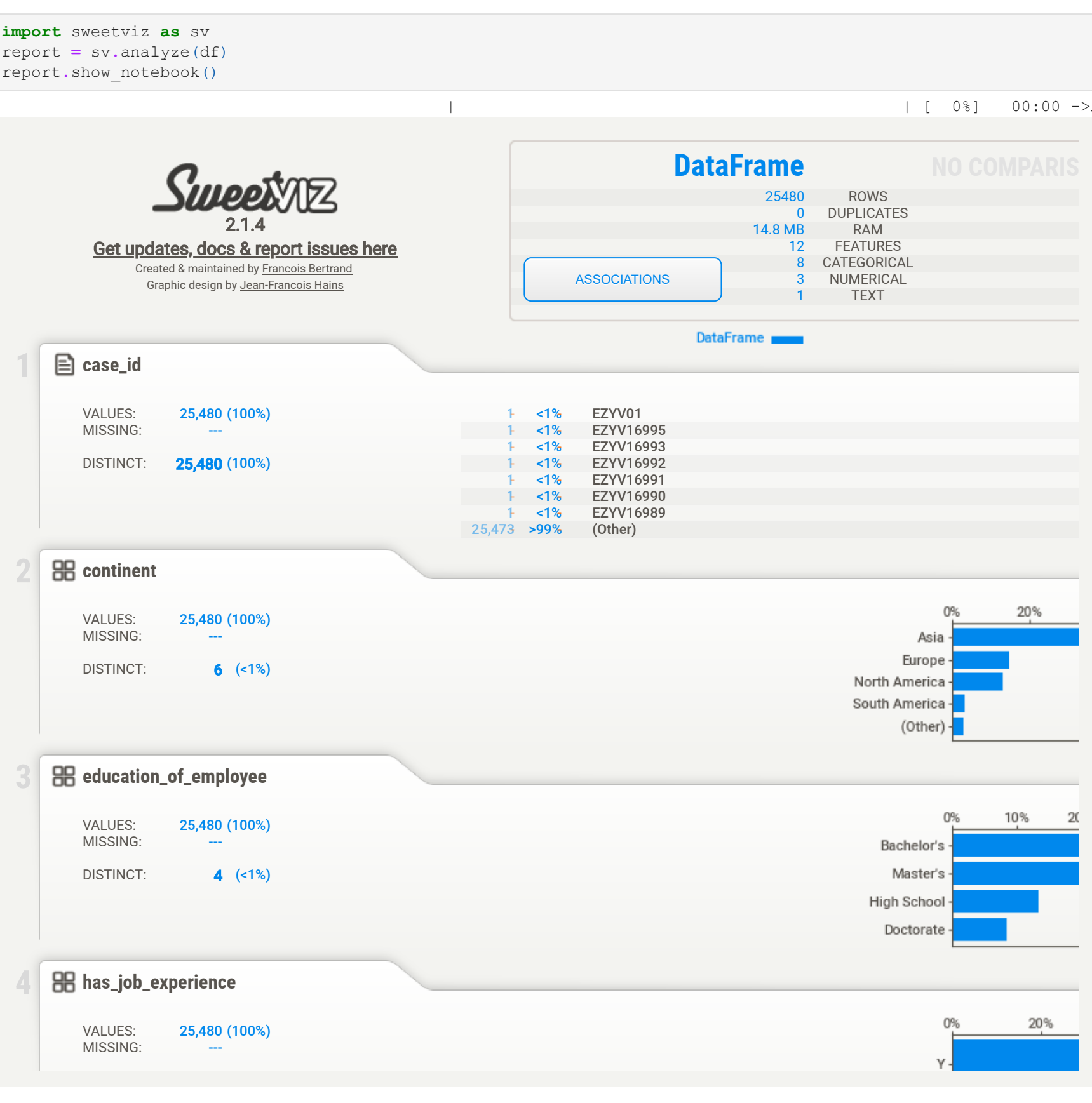
SweetViz is a python library that generates beautiful, high-density visualizations which is useful for EDA with just ffew lines of code.

The SweetViz library generates a report having:

- An overview of the dataset.
- Variable properties.
- Categorical associations.
- Numerical associations.
- Most frequent, smallest, largest values for numerical features.

In [16]:
```
import sweetviz as sv
report = sv.analyze(df)
report.show_notebook()
```

| | [0%]   00:00 ->...



3. AutoViz Tools

AutoViz can find the most important features and plot impactful visualizations only using those automatically selected features.

The Autoviz library generates a report having:

- An overview of the dataset.
- Pairwise scatter plot of continuous variables.
- Distribution of categorical variables.
- Heatmaps of continuous variables.
- Average numerical variable by each categorical variable.

In [17]:
```
pip install xlrd
```

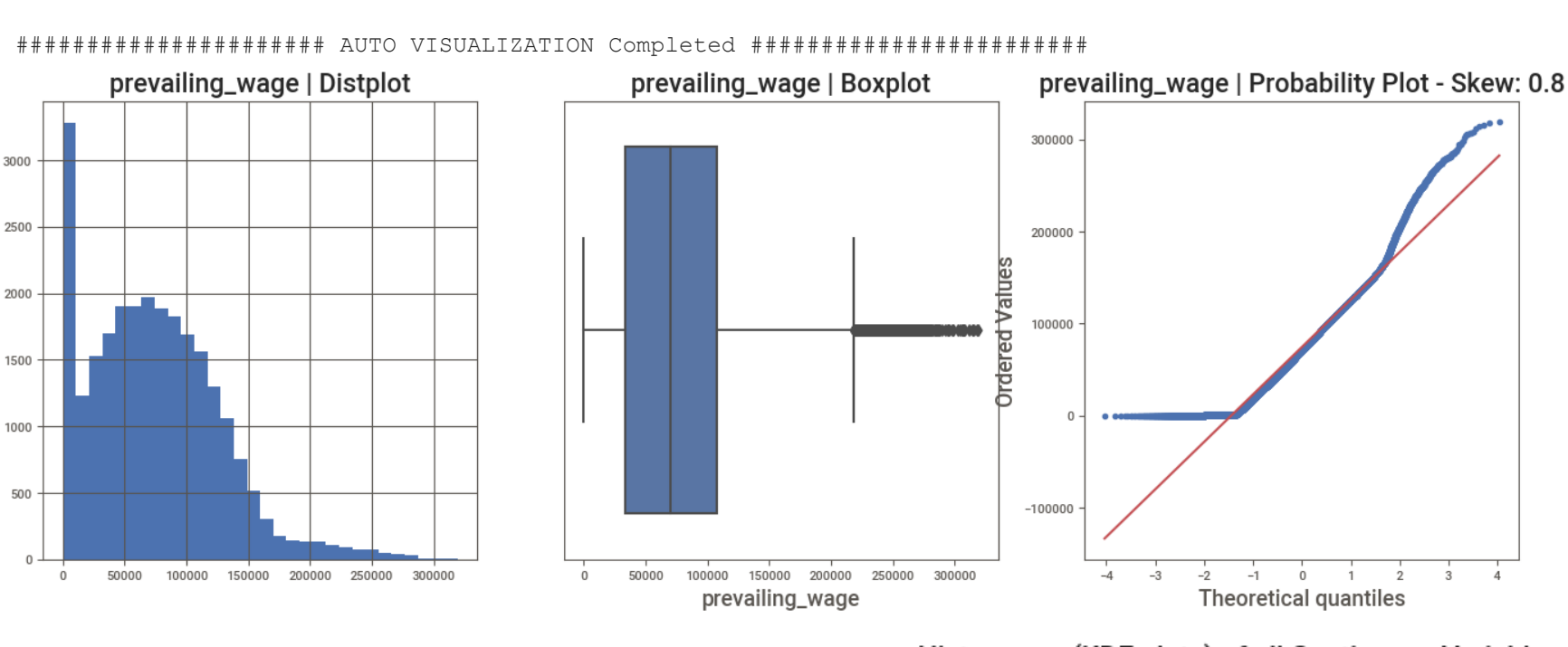
```
Requirement already satisfied: xlrd in c:\users\nilut\anaconda3\lib\site-packages (2.0.1)
Note: you may need to restart the kernel to use updated packages.
```

In [18]:
```
from autoviz.AutoViz_Class import AutoViz_Class
av = AutoViz_Class()
df_av = av.AutoViz('D:\Data Science\Dataset\data5\Visadataset.csv')
```

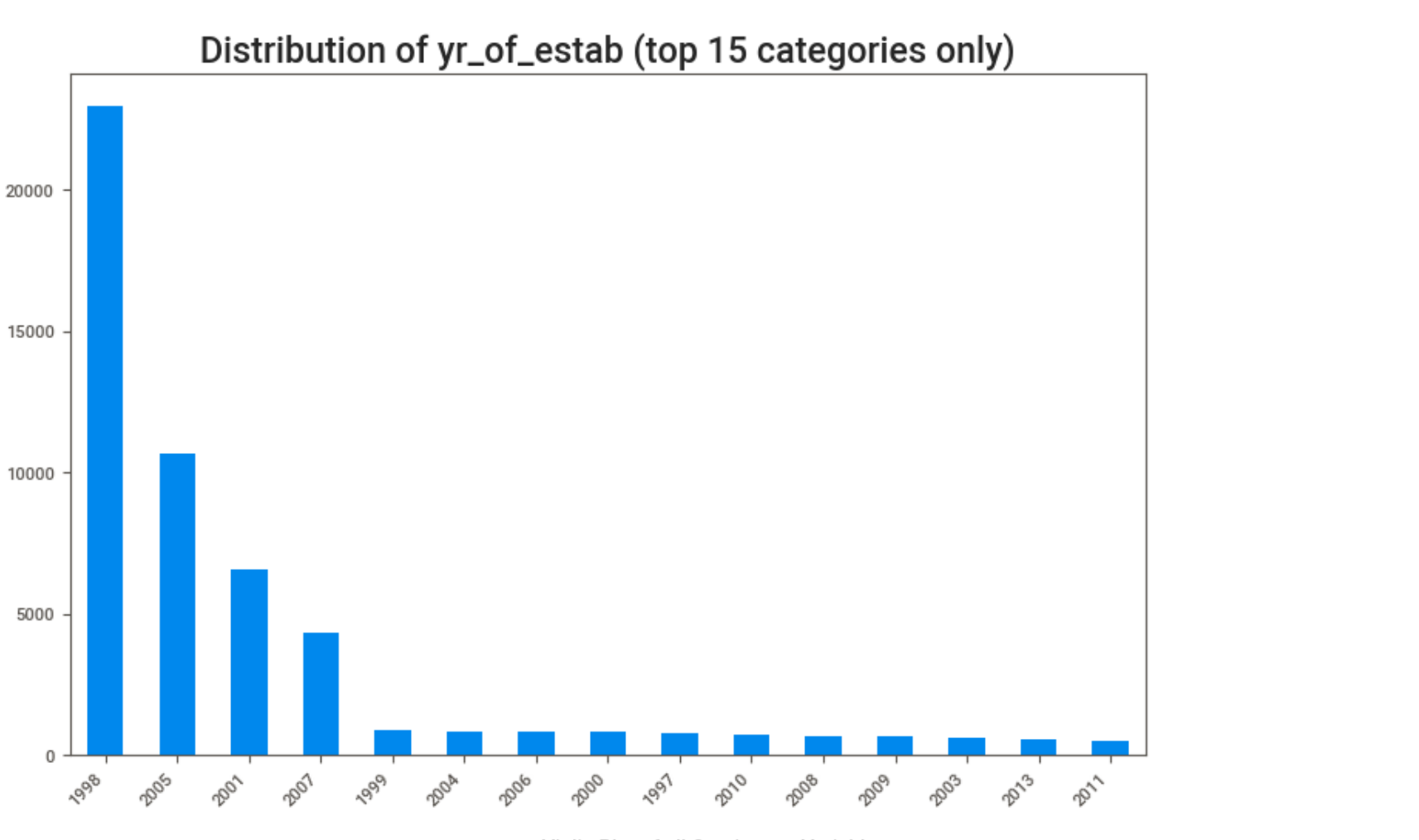
```
Shape of your Data Set loaded: (25480, 12)
#######################################################################################
######################## C L A S S I F Y I N G   V A R I A B L E S  ####################
#######################################################################################
Classifying variables in data set...
```

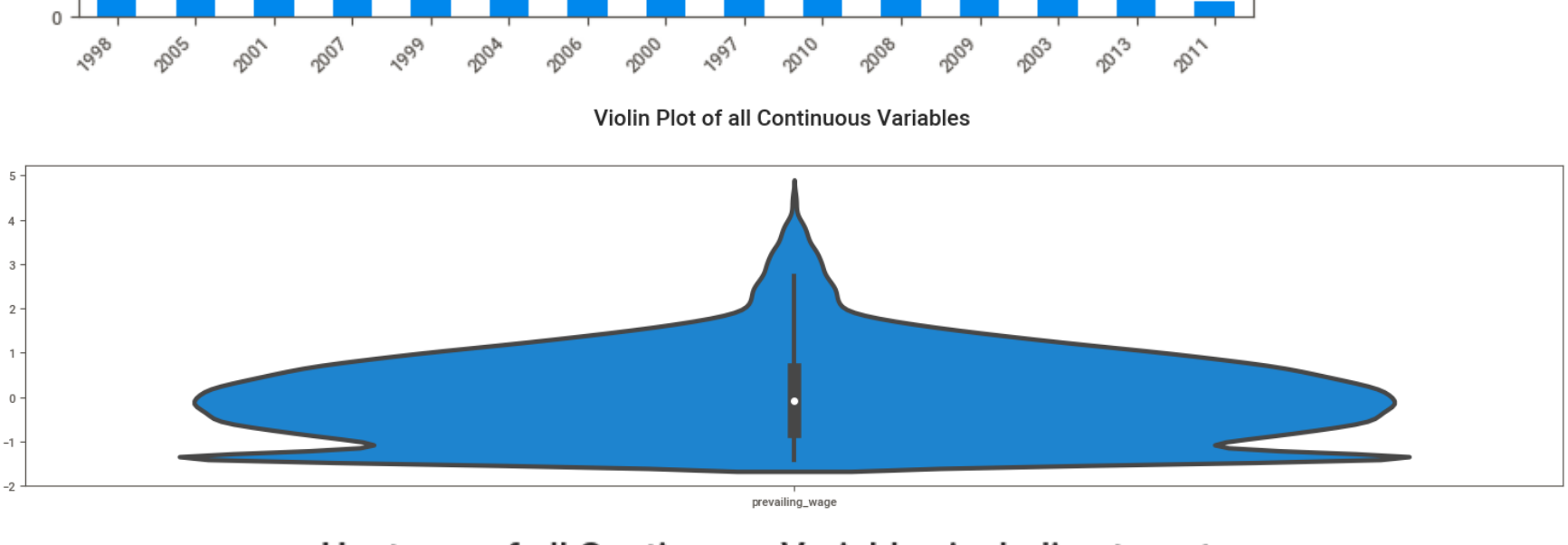
	Nuniques	dtype	Nulls	Nullpercent	NuniquePercent	Value counts Min	Data cleaning improvement suggestions
case_id	25480	object	0	0.000000	100.000000	1	combine rare categories, possible ID column: drop
prevailing_wage	25454	float64	0	0.000000	99.897959	0	
no_of_employees	7105	int64	0	0.000000	27.884615	0	
yr_of_estab	199	int64	0	0.000000	0.781005	0	
continent	6	object	0	0.000000	0.023548	192	
region_of_employment	5	object	0	0.000000	0.019623	375	
education_of_employee	4	object	0	0.000000	0.015699	2192	
unit_of_wage	4	object	0	0.000000	0.015699	89	
has_job_experience	2	object	0	0.000000	0.007849	10678	
requires_job_training	2	object	0	0.000000	0.007849	2955	
full_time_position	2	object	0	0.000000	0.007849	2707	
case_status	2	object	0	0.000000	0.007849	8462	

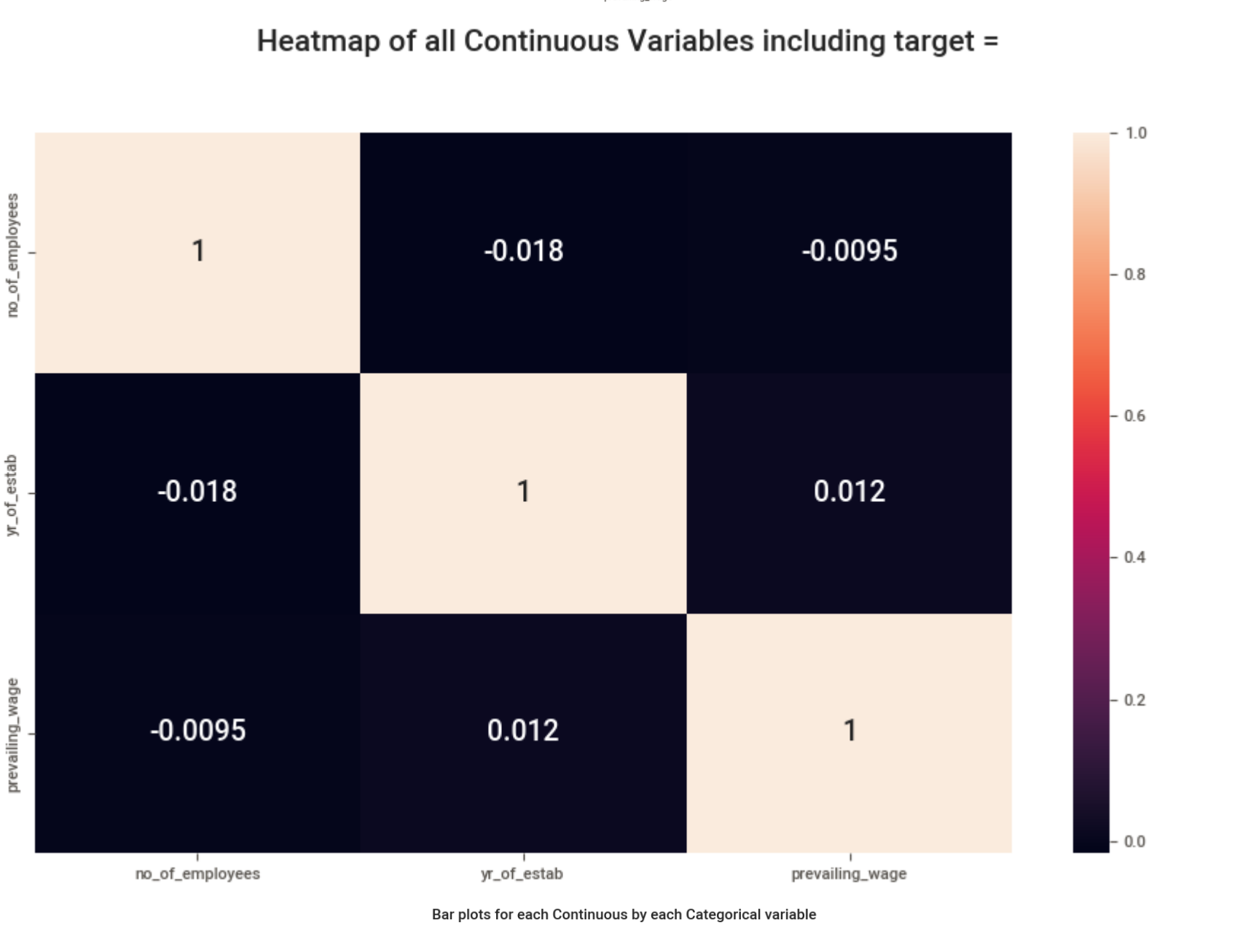
```
    12 Predictors classified...
        1 variables removed since they were ID or low-information variables
        List of variables removed: ['case_id']
All Plots done
Time to run AutoViz = 1 seconds

###################### AUTO VISUALIZATION Completed ########################
```

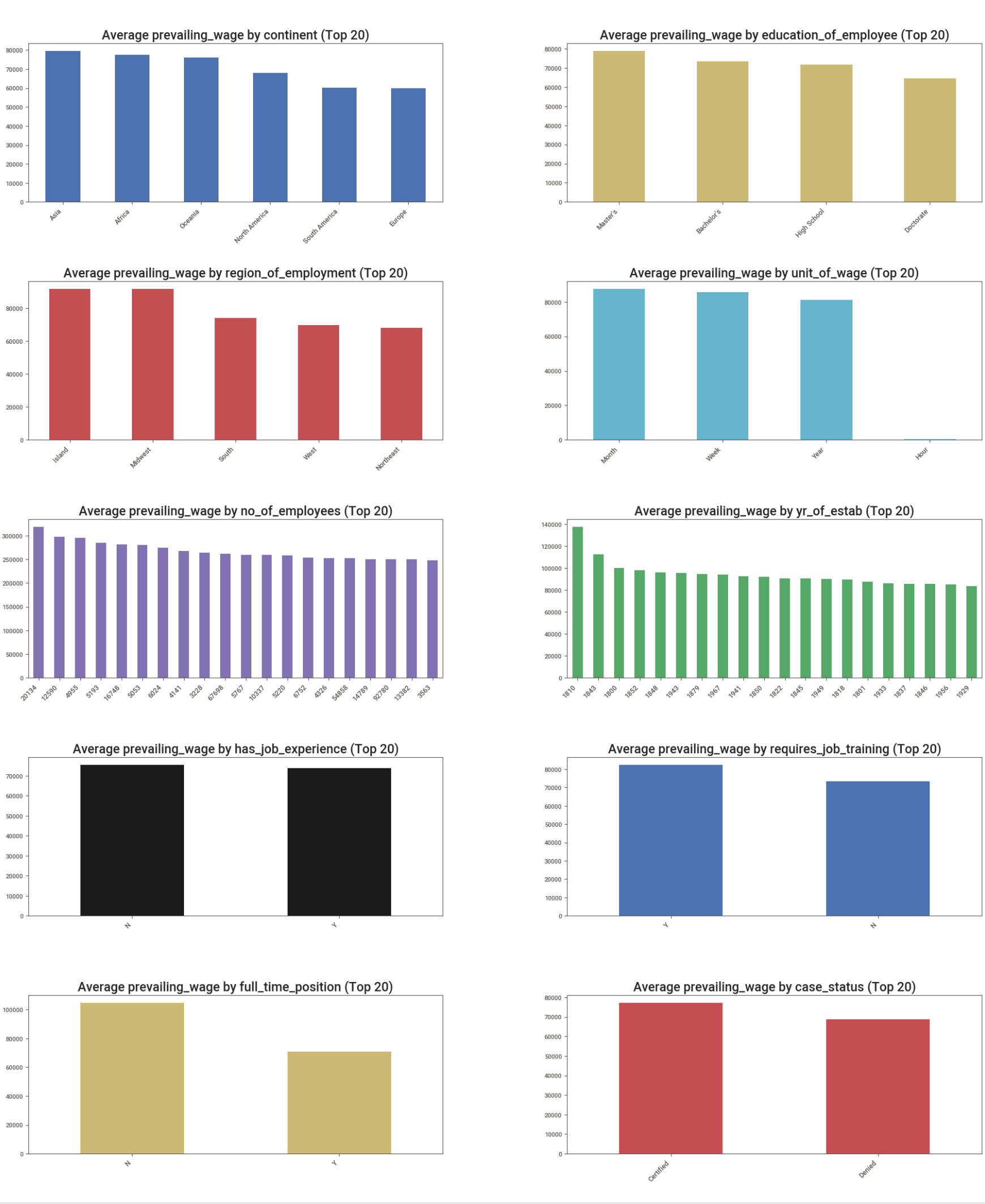


Histograms (KDE plots) of all Continuous Variables









In [20]:
```
pip install xlrd
```

```
Requirement already satisfied: xlrd in c:\users\nilut\anaconda3\lib\site-packages (2.0.1)
Note: you may need to restart the kernel to use updated packages.
```

4. DABL Tools

Data Analysis Baseline library - Dabl. It is a python library providing a quick overview via visualizations, as well as convenient preprocessing and model search for machine learning.

The DABL library generates a report having:

- Bar plot for target distribution
- Scatter Pair plots
- Linear Discriminant Analysis

In [25]:
```
import pandas as pd
import dabl
dabl.plot(df, target_col="case_status")
```

```
Target looks like classification
Linear Discriminant Analysis training set score: 0.500
```

Out[25]:
```
[[<Figure size 900x900 with 12 Axes>,
  <Figure size 1600x400 with 4 Axes>,
  <Figure size 800x550 with 1 Axes>],
 None]
```