

EDA Implementation with the help of Automated EDA Tools:-

Submitted By :- Ambarish Singh

```
In [1]: #Importing CSV Dataset with a help of Pandas Library:-  
  
import pandas as pd  
df = pd.read_csv('Visadataset.csv')
```

```
In [2]: # head() used to provide top 5 data from dataset.  
df.head()
```

```
Out[2]:
```

	case_id	continent	education_of_employee	has_job_experience	requires_job_training	no_of_employee
0	EZYV01	Asia	High School	N	N	1451
1	EZYV02	Asia	Master's	Y	N	241
2	EZYV03	Asia	Bachelor's	N	Y	4444
3	EZYV04	Asia	Bachelor's	N	N	9
4	EZYV05	Africa	Master's	Y	N	108

Performing EDA with the help of Pandas-Profiling Automated EDA Tools.

Pandas Profiling Tools

--> It is an open-source library written in python and it generated interactive HTML reports and describes various aspects of the dataset. Key functionalities include handling missing values, statistics of dataset like mean, mode, median, skewness, standard deviation etc, charts like histograms and correlations as well.

```
In [3]: ## Performing EDA with the help of Pandas- Profiling Automated EDA Tools.  
  
import pandas as pd  
from pandas_profiling import ProfileReport  
profile = ProfileReport(df, title="Report")  
profile
```

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	2.3 MiB
Average record size in memory	96.0 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
case_id has unique values	

Out[3]:

Performing EDA with the help of Sweetviz Automated EDA Tools.

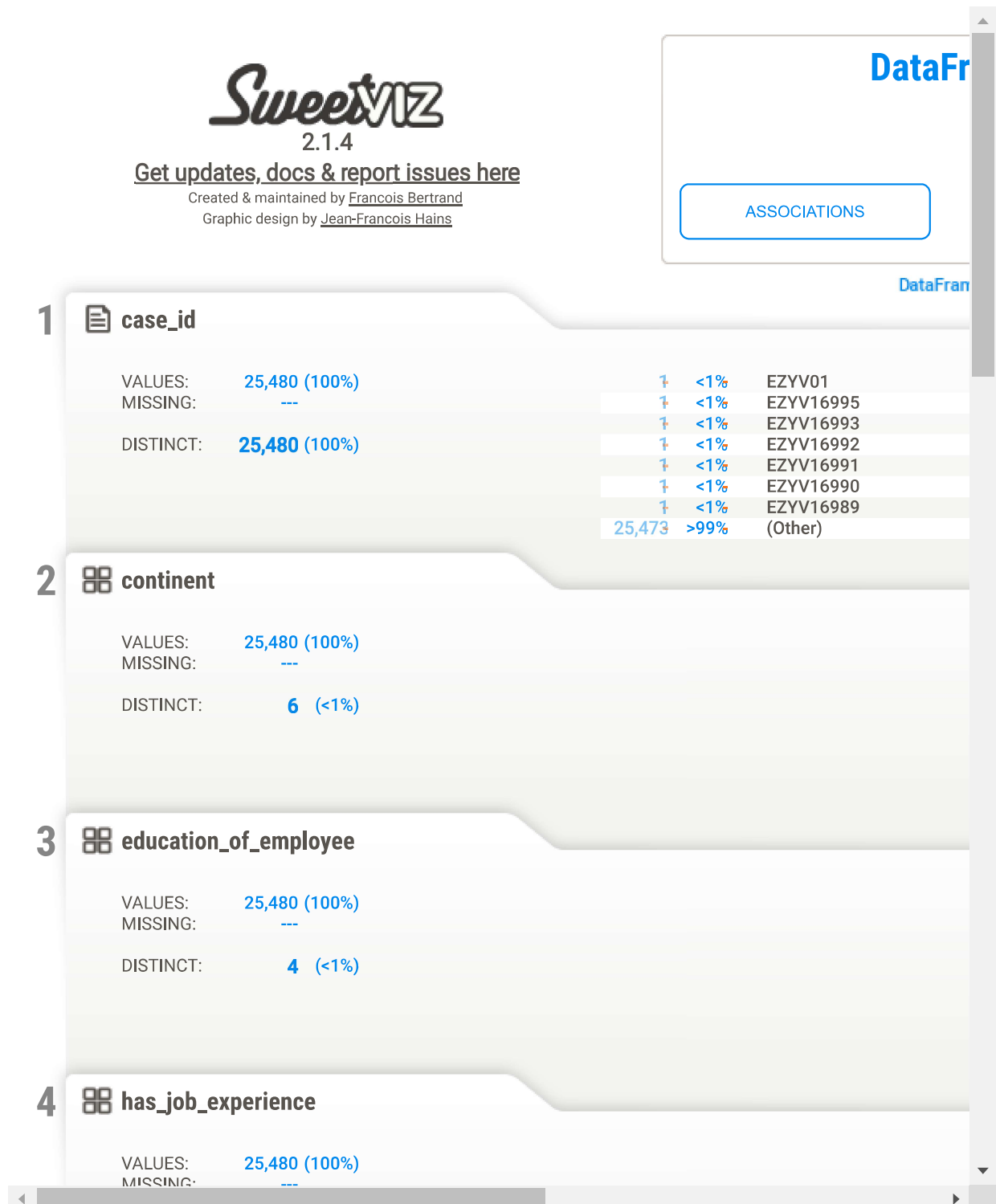
SweetViz Tools

--> It is an open-source python library that used to get visualizations which is useful in exploratory data analysis with just a few lines of codes. The library can be used to visualize the variables and comparing the dataset.

In [34]:

```
## Performing EDA with the help of Sweet viz Automated EDA Tools.
```

```
import sweetviz as sv
analyze_report = sv.analyze(df)
analyze_report.show_notebook()
```



Performing EDA with the help of Autoviz Automated EDA Tools.

Autoviz Tools

-->It stands for Automatically Visualize. Visualization is possible with any size of the dataset with a few lines of code.

In [35]:

```
## Performing EDA with the help of Autoviz Automated EDA Tools.
```

```
from autoviz.AutoViz_Class import AutoViz_Class
AV = AutoViz_Class()
df_av = AV.AutoViz('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...

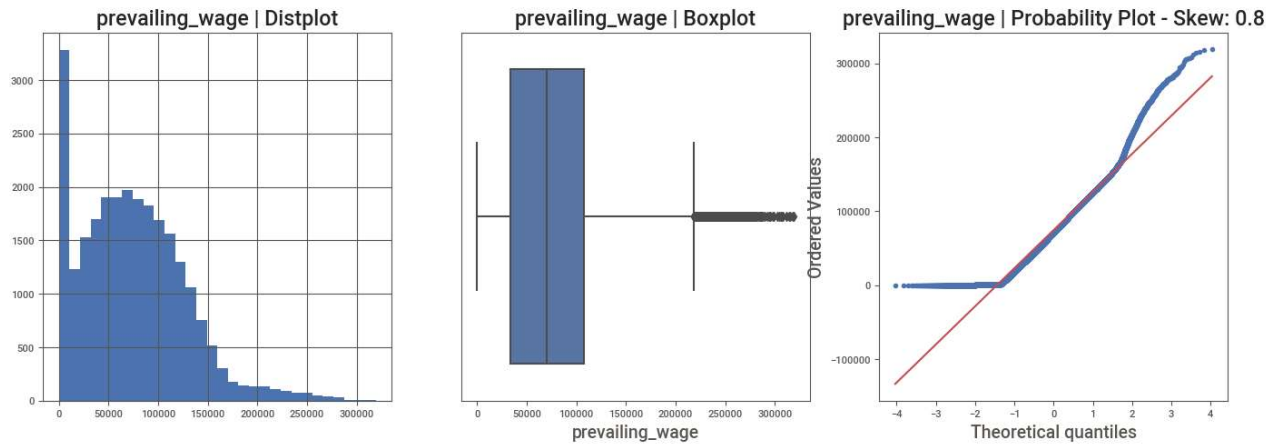
1 variables removed since they were ID or low-information variables

List of variables removed: ['case_id']

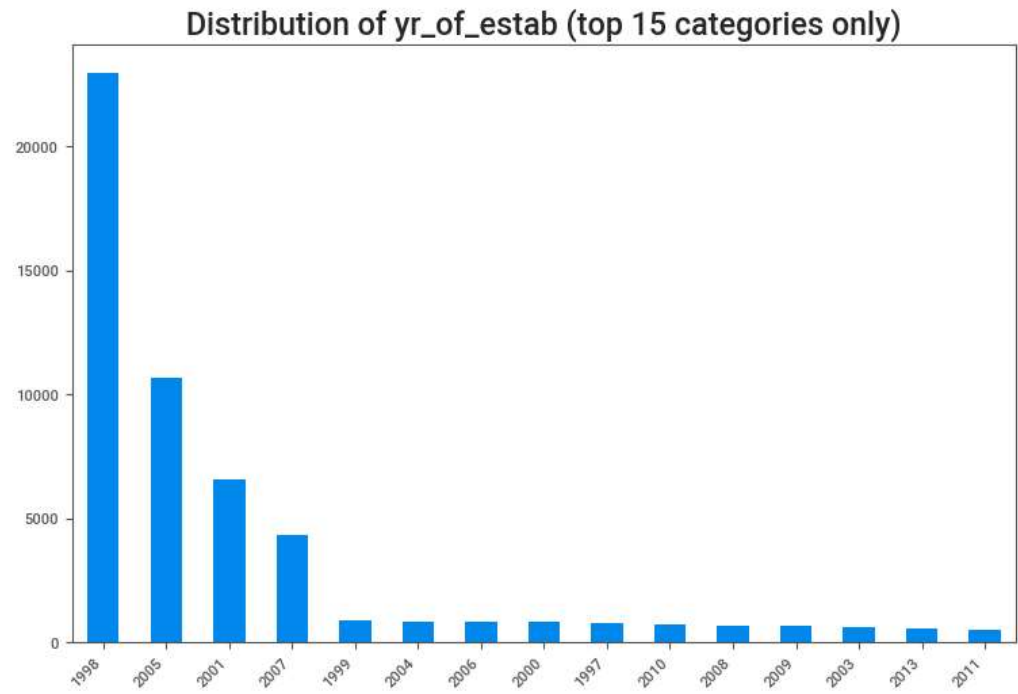
All Plots done

Time to run AutoViz = 3 seconds

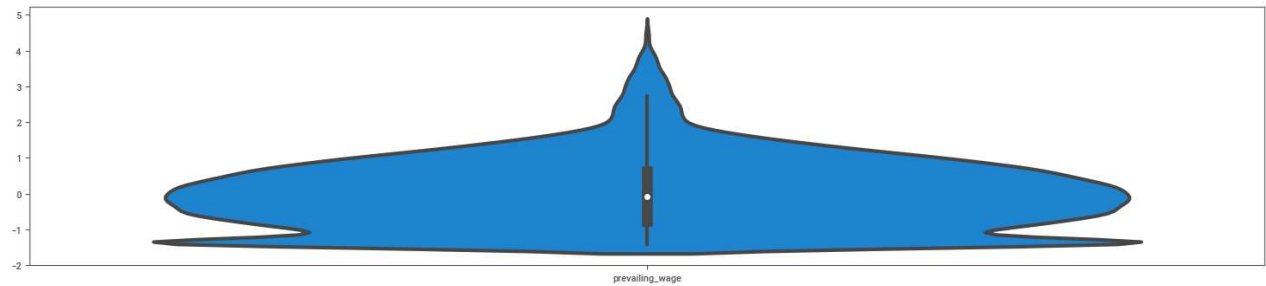
```
##### AUTO VISUALIZATION Completed #####
```



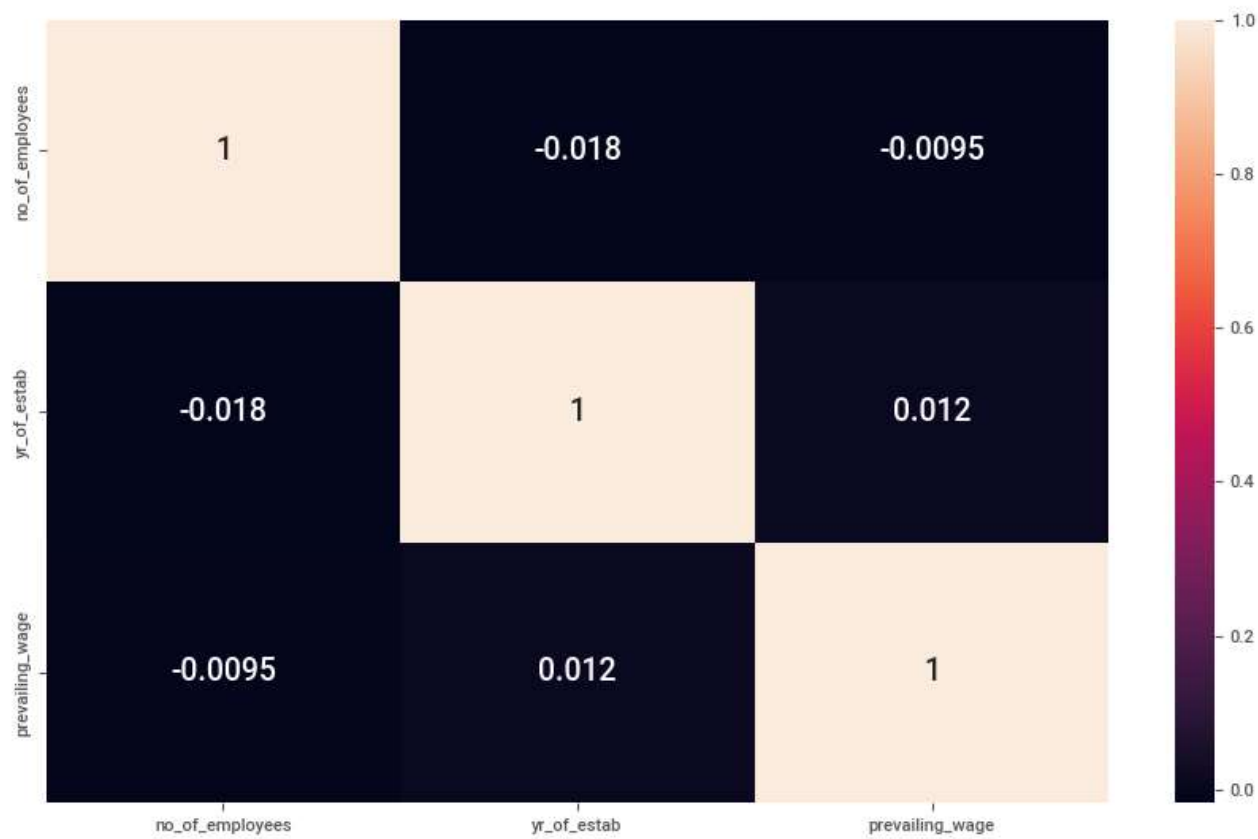
Histograms (KDE plots) of all Continuous Variables



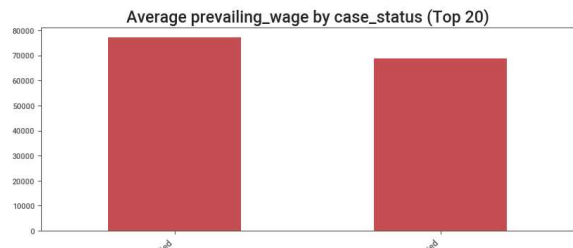
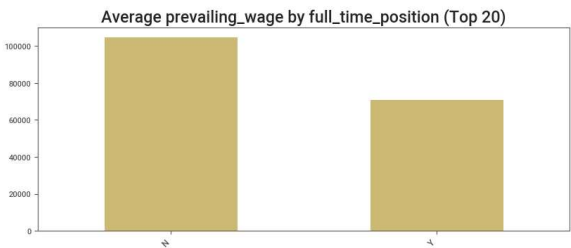
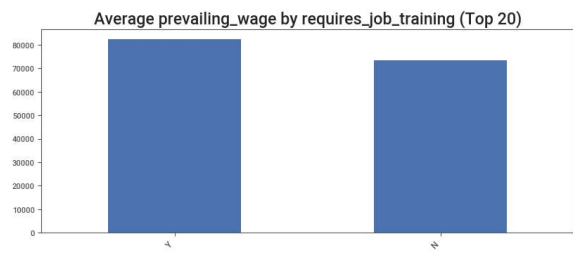
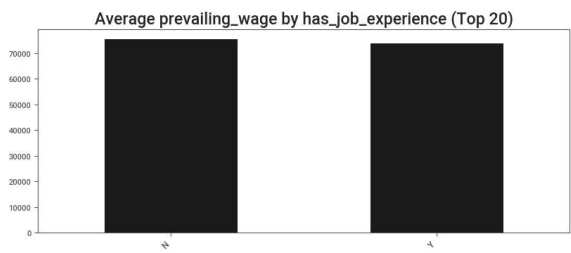
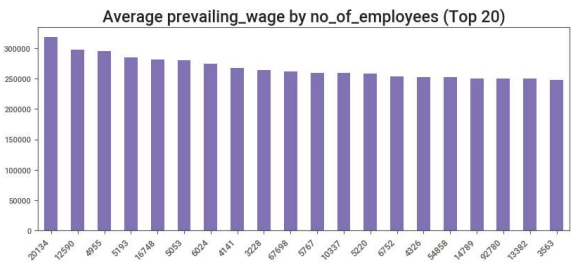
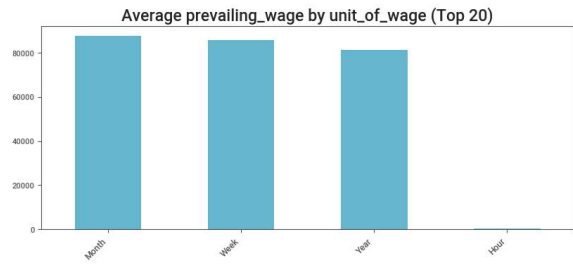
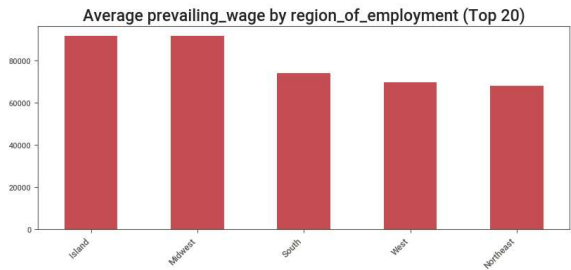
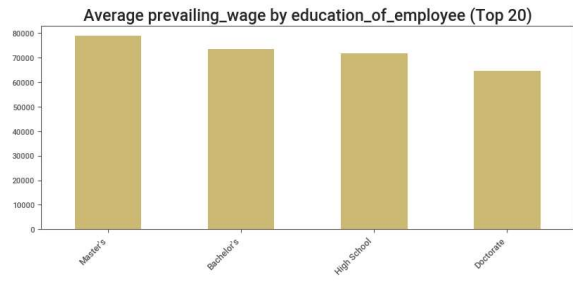
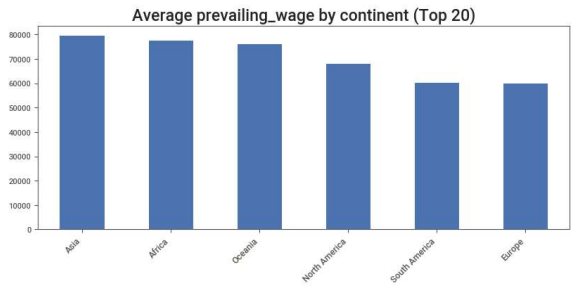
Violin Plot of all Continuous Variables



Heatmap of all Continuous Variables including target =



Bar plots for each Continuous by each Categorical variable



Thank You