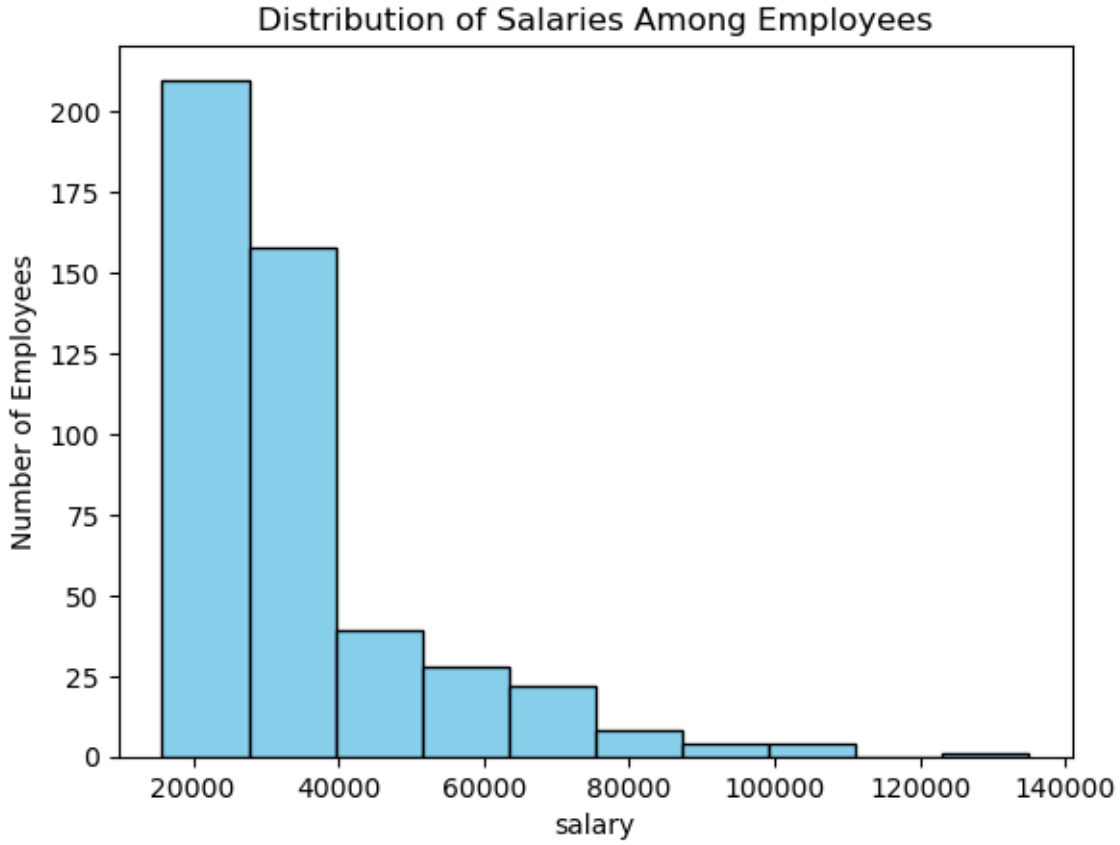
Q.1. Generate a histogram using any library to visualize the distribution of salaries among employees in the dataset.

Ans:

```
import pandas as pd

import matplotlib.pyplot as plt
sal = pd.read_csv(r'D:\IBMData\Employee data.csv')

salary_data = df['salary']

# Create a histogram

plt.hist(salary_data, bins=10, color='skyblue', edgecolor='black')

# Add labels and title

plt.xlabel('salary')

plt.ylabel('Number of Employees')

plt.title('Distribution of Salaries Among Employees')

# Show the plot

plt.show()
```

output:



Q.2. Generate a bar plot to compare the average salary of male and female employees using any library

Ans:

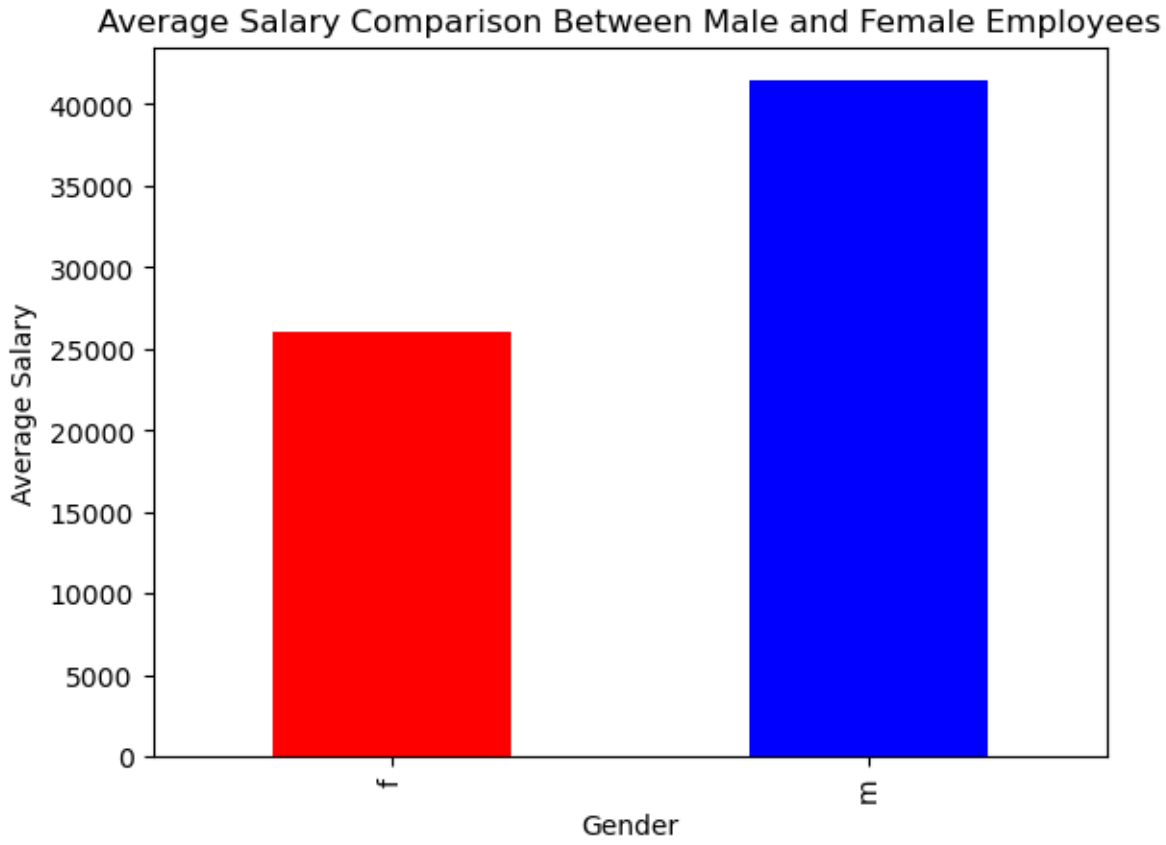
```
import pandas as pd

import matplotlib.pyplot as plt
sal = pd.read_csv(r'D:\IBMData\Employee data.csv')

grouped_data = df.groupby('gender')['salary'].mean()

# Create a bar plot

grouped_data.plot(kind='bar', color=['red', 'blue'])

# Add labels and title

plt.xlabel('Gender')

plt.ylabel('Average Salary')
```

plt.title('Average Salary Comparison Between Male and Female Employees')

Show the plot

plt.show()

output:



Q.3. Create a scatter plot using any library to illustrate the relationship between previous work experience (prevexp) and the current salary of employees.

Ans:

import pandas as pd

import matplotlib.pyplot as plt

Load the CSV file into a DataFrame

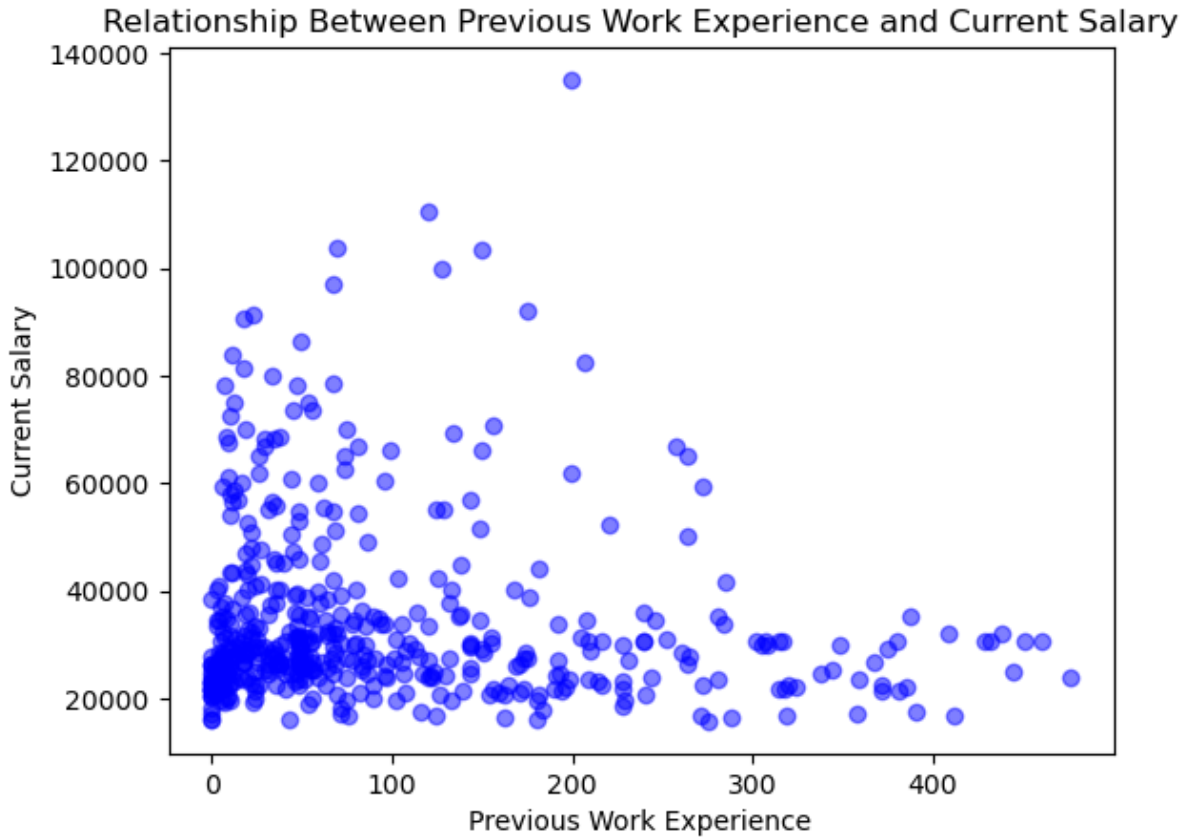
sal = pd.read_csv(r'D:\IBMData\Employee data.csv')

plt.scatter(sal['prevexp'], sal['salary'], color='blue', alpha=0.5)

Add labels and title

```
plt.xlabel('Previous Work Experience')

plt.ylabel('Current Salary')

plt.title('Relationship Between Previous Work Experience and Current Salary')

# Show the plot

plt.show()
```

output:



Q.4. Generate a pie chart to visualize the distribution of educational backgrounds among employees.

Ans:

```
import pandas as pd

import matplotlib.pyplot as plt

# Load the CSV file into a DataFrame

sal = pd.read_csv(r'D:\IBMData\Employee data.csv')

education_counts = sal['educ'].value_counts()
```

Create a pie chart

plt.pie(education_counts, labels=education_counts.index, autopct='%1.1f%%', startangle=90, colors=['orange', 'lightgreen', 'pink', 'yellow'])

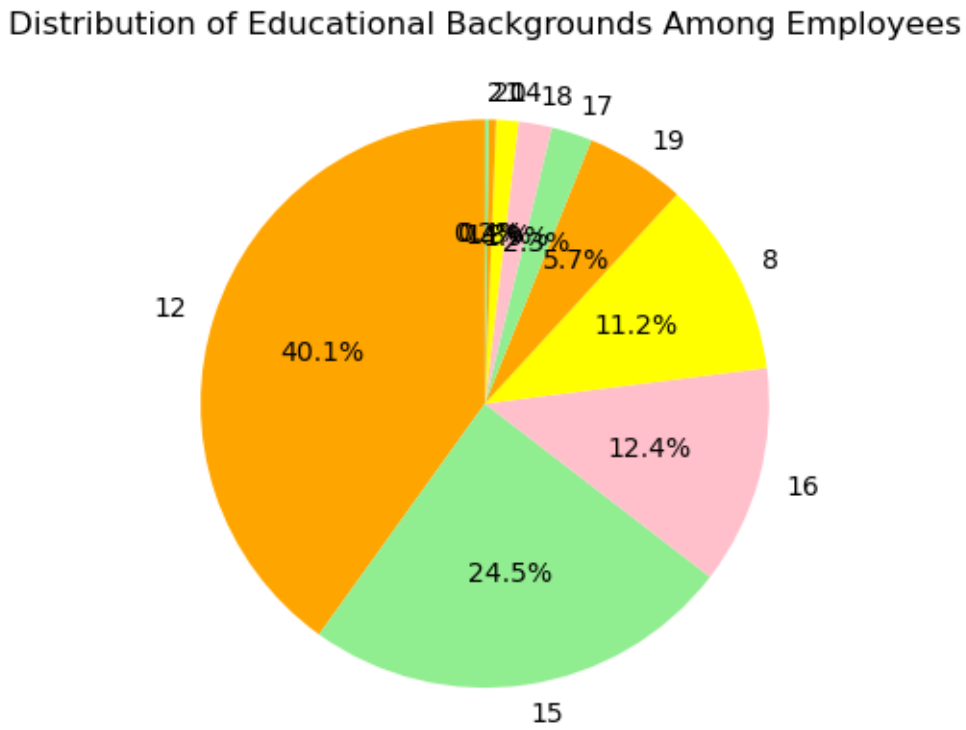
Add title

plt.title('Distribution of Educational Backgrounds Among Employees')

Show the plot

plt.show()

output:



Q.5. Use any AI library such as sweetviz or dtale to generate the summary of the data.

Ans:

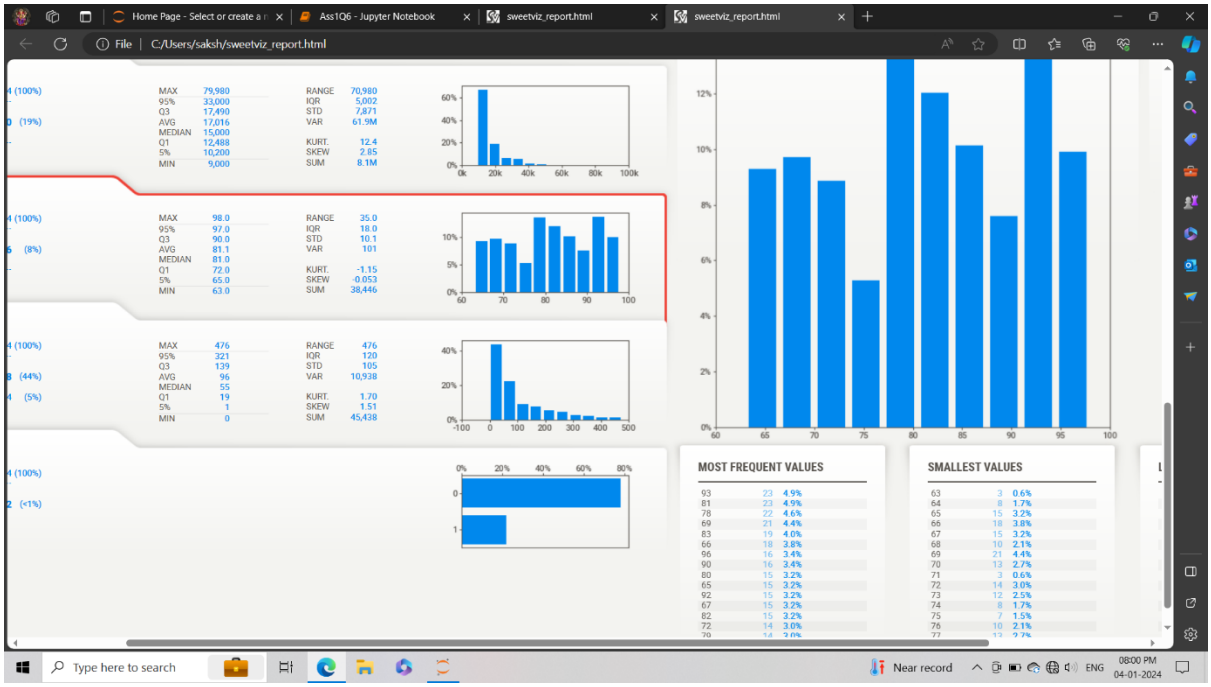
Using sweetviz library

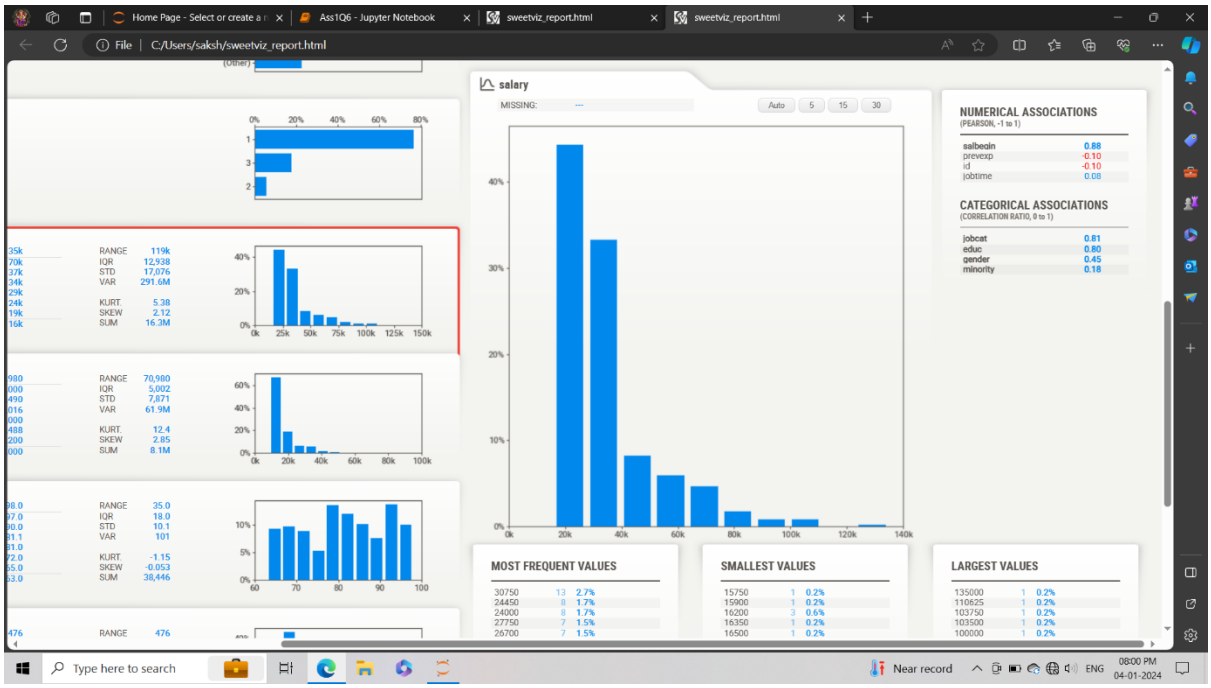
import pandas as pd

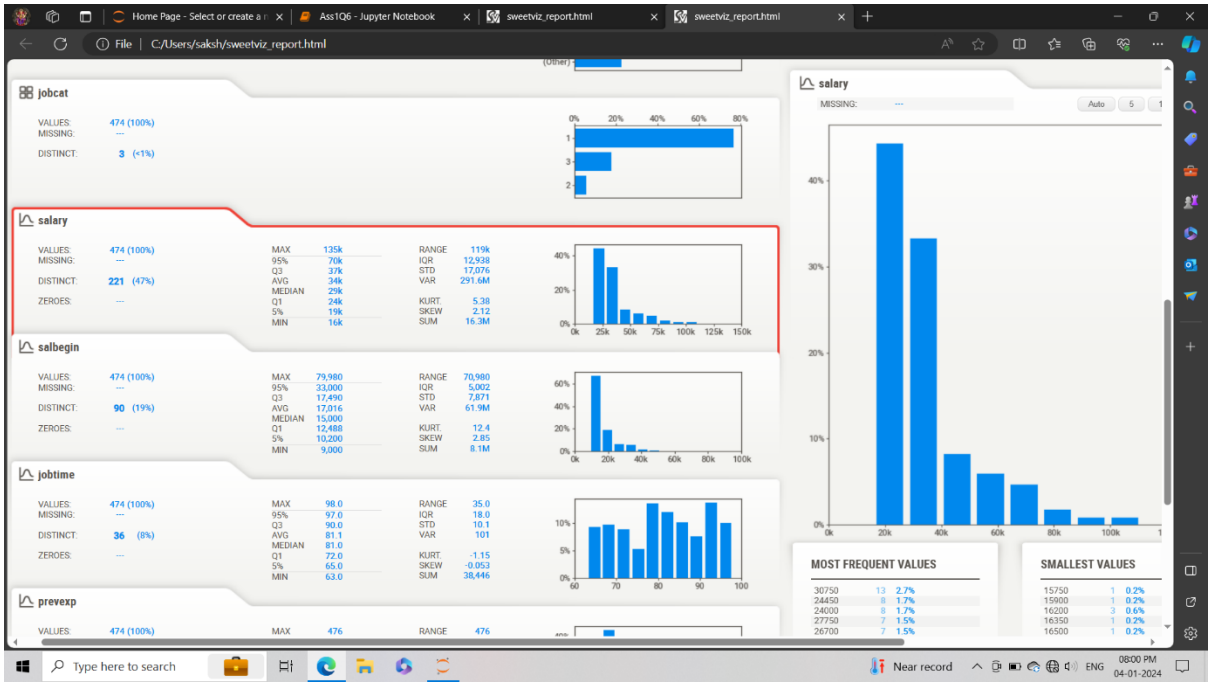
import sweetviz as sv

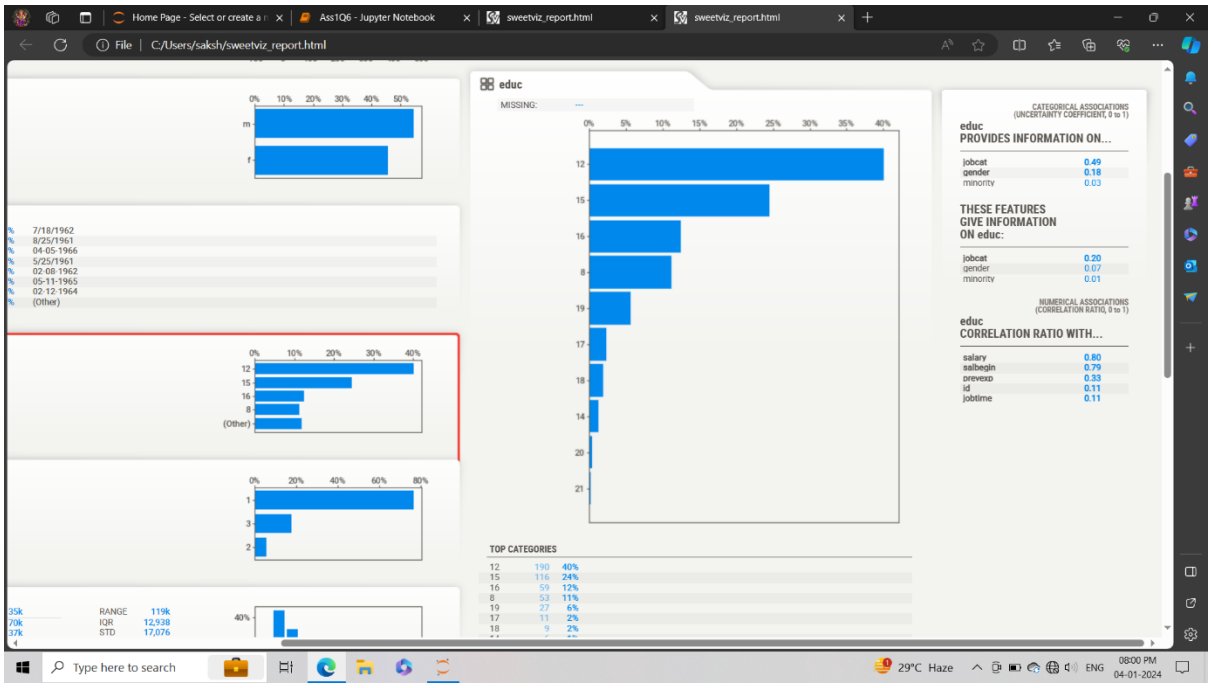
Load the CSV file into a DataFrame

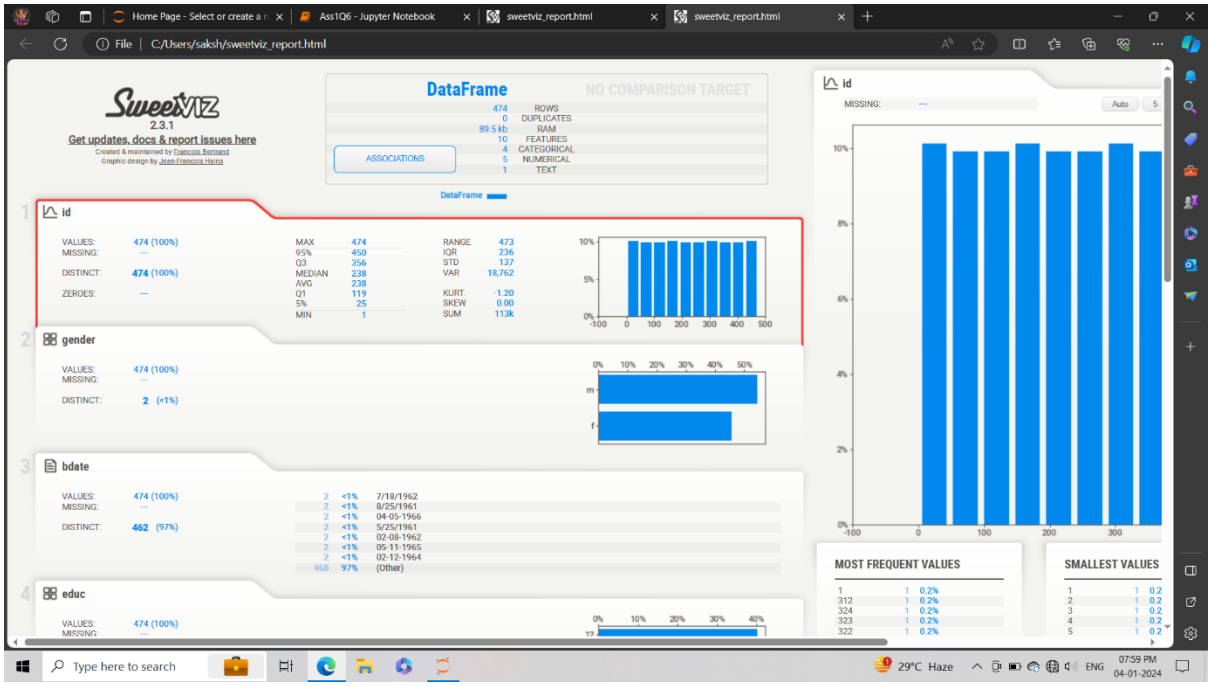
```
sal = pd.read_csv(r'D:\IBMData\Employee data.csv')

# Generate the data summary report

data_report = sv.analyze(sal)

# Save the report to an HTML file

data_report.show_html('sweetviz_report.html')
```

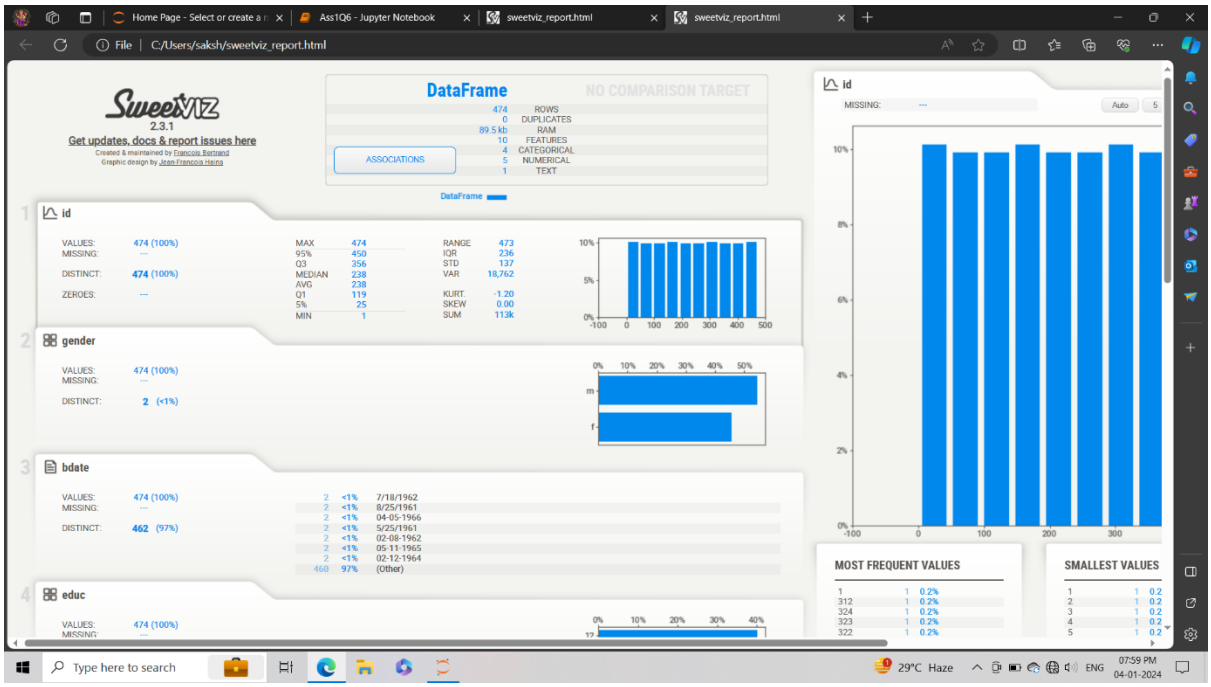





Home Page - Select or create a r
Ass1Q6 - Jupyter Notebook
sweetviz_report.html
sweetviz_report.html
File | C:/Users/saksh/sweetviz_report.html
jobcat
VALUES: 474 (100%)
MISSING: ---
DISTINCT: 3 (<1%)
salary
VALUES: 474 (100%)
MISSING: ---
DISTINCT: 221 (47%)
ZEROES: ---
MAX 135k
95% 70k
Q3 37k
AVG 34k
MEDIAN 29k
Q1 24k
5% 19k
MIN 16k
RANGE 119k
IQR 12,938
STD 17,076
VAR 291.6M
KURT 5.38
SKEW 2.12
SUM 16.3M
salbegin
VALUES: 474 (100%)
MISSING: ---
DISTINCT: 90 (19%)
ZEROES: ---
MAX 79,980
95% 33,000
Q3 17,490
AVG 17,016
MEDIAN 15,000
Q1 12,488
5% 10,200
MIN 9,000
RANGE 70,980
IQR 5,002
STD 7,871
VAR 61.9M
KURT 12.4
SKEW 2.85
SUM 8.1M
jobtime
VALUES: 474 (100%)
MISSING: ---
DISTINCT: 36 (8%)
ZEROES: ---
MAX 98.0
95% 97.0
Q3 90.0
AVG 81.1
MEDIAN 81.0
Q1 72.0
5% 65.0
MIN 63.0
RANGE 35.0
IQR 18.0
STD 10.1
VAR 101
KURT -1.15
SKEW -0.053
SUM 38,446
prevexp
VALUES: 474 (100%)
MAX 476
RANGE 476
salary
MISSING: ---
MOST FREQUENT VALUES
30750 13 2.7%
24450 8 1.7%
24000 8 1.7%
27750 7 1.5%
26700 7 1.5%
SMALLEST VALUES
15750 1 0.2%
15900 1 0.2%
16200 3 0.6%
16350 1 0.2%
16500 1 0.2%
Type here to search
Near record
ENG
08:00 PM
04-01-2024


Home Page - Select or create a r
Ass1Q6 - Jupyter Notebook
sweetviz_report.html
sweetviz_report.html
File | C:/Users/saksh/sweetviz_report.html
7/18/1962
8/25/1961
04-05-1966
5/25/1961
02-08-1962
05-11-1965
02-12-1964
(Other)
RANGE 119k
IQR 12,938
STD 17,076
educ
MISSING: ---
CATEGORICAL ASSOCIATIONS (UNCERTAINTY COEFFICIENT, 0 to 1)
educ
PROVIDES INFORMATION ON...
jobcat 0.49
gender 0.18
minority 0.03
THESE FEATURES GIVE INFORMATION ON educ:
jobcat 0.20
gender 0.07
minority 0.01
NUMERICAL ASSOCIATIONS (CORRELATION RATIO, 0 to 1)
educ
CORRELATION RATIO WITH...
salary 0.80
salbegin 0.79
prevexp 0.33
id 0.11
jobtime 0.11
TOP CATEGORIES
12 190 40%
15 116 24%
16 59 12%
8 53 11%
19 27 6%
17 11 2%
18 9 2%
Type here to search
29°C Haze
ENG
08:00 PM
04-01-2024


Home Page - Select or create a r
Ass1Q6 - Jupyter Notebook
sweetviz_report.html
sweetviz_report.html
File | C:/Users/saksh/sweetviz_report.html
SweetViz
2.3.1
Get updates, docs & report issues here
Created & maintained by Francois Bertrand
Graphic design by Jean-Francois Haina
DataFrame
NO COMPARISON TARGET
474 ROWS
0 DUPLICATES
89.5 kb RAM
10 FEATURES
4 CATEGORICAL
5 NUMERICAL
1 TEXT
ASSOCIATIONS
DataFrame
1 id
VALUES: 474 (100%)
MISSING: ---
DISTINCT: 474 (100%)
ZEROES: ---
MAX 474
95% 450
Q3 356
MEDIAN 238
AVG 238
Q1 119
5% 25
MIN 1
RANGE 473
IQR 236
STD 137
VAR 18,762
KURT -1.20
SKEW 0.00
SUM 113k
2 gender
VALUES: 474 (100%)
MISSING: ---
DISTINCT: 2 (<1%)
3 bdate
VALUES: 474 (100%)
MISSING: ---
DISTINCT: 462 (97%)
2 <1% 7/18/1962
2 <1% 8/25/1961
2 <1% 04-05-1966
2 <1% 5/25/1961
2 <1% 02-08-1962
2 <1% 05-11-1965
2 <1% 02-12-1964
460 97% (Other)
4 educ
VALUES: 474 (100%)
MISSING: ---
id
MISSING: ---
Auto
MOST FREQUENT VALUES
SMALLEST VALUES
Type here to search
29°C Haze
ENG
07:59 PM
04-01-2024


Home Page - Select or create a r
Ass1Q6 - Jupyter Notebook
sweetviz_report.html
sweetviz_report.html
File | C:/Users/saksh/sweetviz_report.html
SweetViz
2.3.1
Get updates, docs & report issues here
Created & maintained by Francois Bertrand
Graphic design by Jean-Francois Haina
DataFrame
NO COMPARISON TARGET
474 ROWS
0 DUPLICATES
89.5 kb RAM
10 FEATURES
4 CATEGORICAL
5 NUMERICAL
1 TEXT
ASSOCIATIONS
DataFrame
1 id
VALUES: 474 (100%)
MISSING: ---
DISTINCT: 474 (100%)
ZEROES: ---
MAX 474
95% 450
Q3 356
MEDIAN 238
AVG 238
Q1 119
5% 25
MIN 1
RANGE 473
IQR 236
STD 137
VAR 18,762
KURT -1.20
SKEW 0.00
SUM 113k
2 gender
VALUES: 474 (100%)
MISSING: ---
DISTINCT: 2 (<1%)
3 bdate
VALUES: 474 (100%)
MISSING: ---
DISTINCT: 462 (97%)
2 <1% 7/18/1962
2 <1% 8/25/1961
2 <1% 04-05-1966
2 <1% 5/25/1961
2 <1% 02-08-1962
2 <1% 05-11-1965
2 <1% 02-12-1964
460 97% (Other)
4 educ
VALUES: 474 (100%)
MISSING: ---
id
MISSING: ---
Auto
MOST FREQUENT VALUES
SMALLEST VALUES
Type here to search
29°C Haze
ENG
07:59 PM
04-01-2024