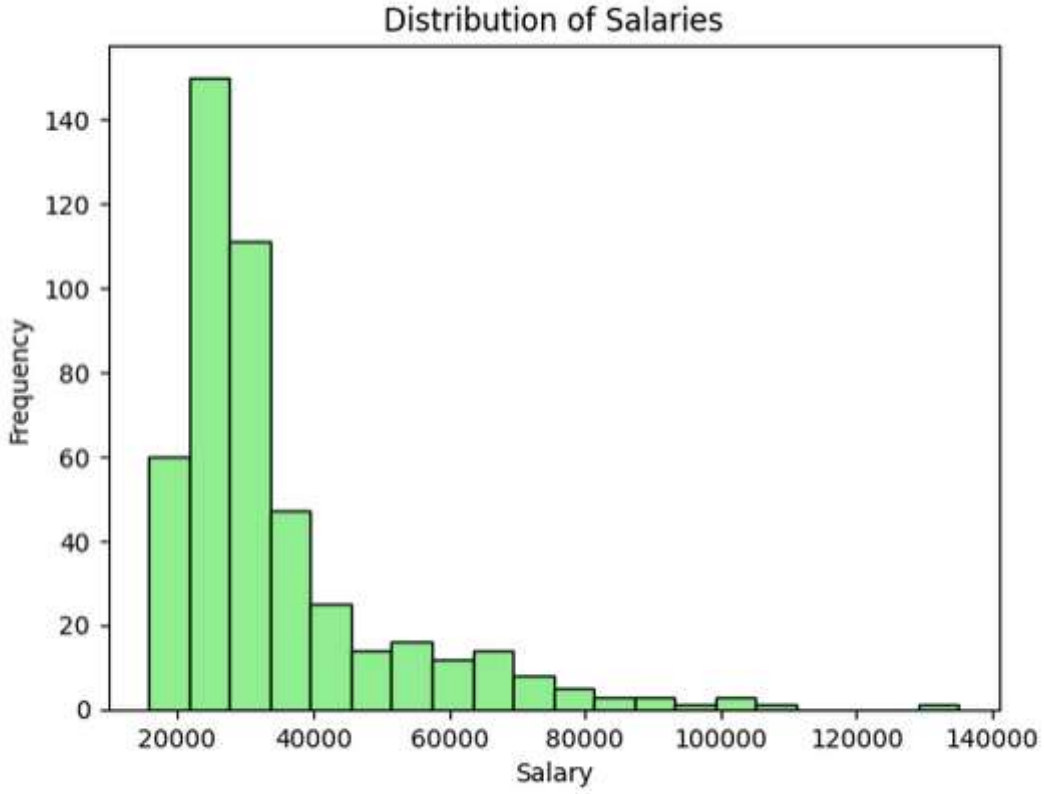
Topic: Data Visualisation using AI

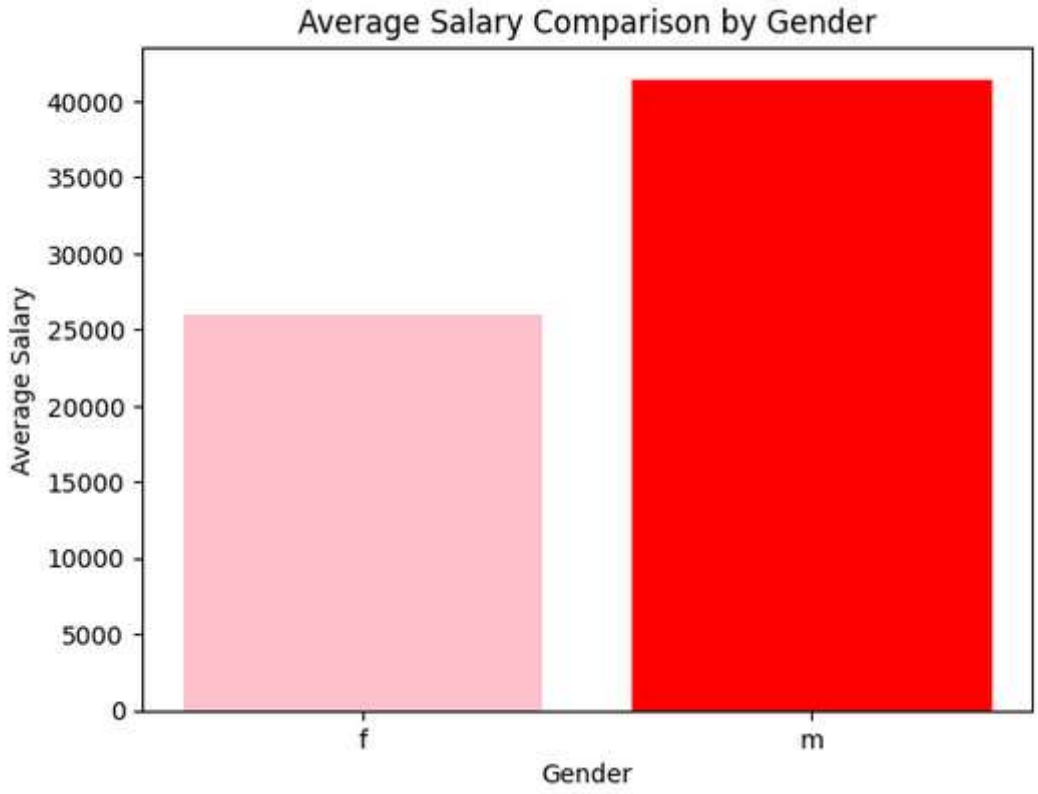
1) Generate a histogram using any library to visualize the distribution of salaries among employees in the dataset.

```
import matplotlib.pyplot as plt
import pandas as pd

df = pd.read_csv(r'C:\Users\dubey\Downloads\Employee data.csv')

plt.hist(df['salary'], bins=20, color='lightgreen', edgecolor='black')

plt.title('Distribution of Salaries')
plt.xlabel('Salary')
plt.ylabel('Frequency')
plt.show()
```



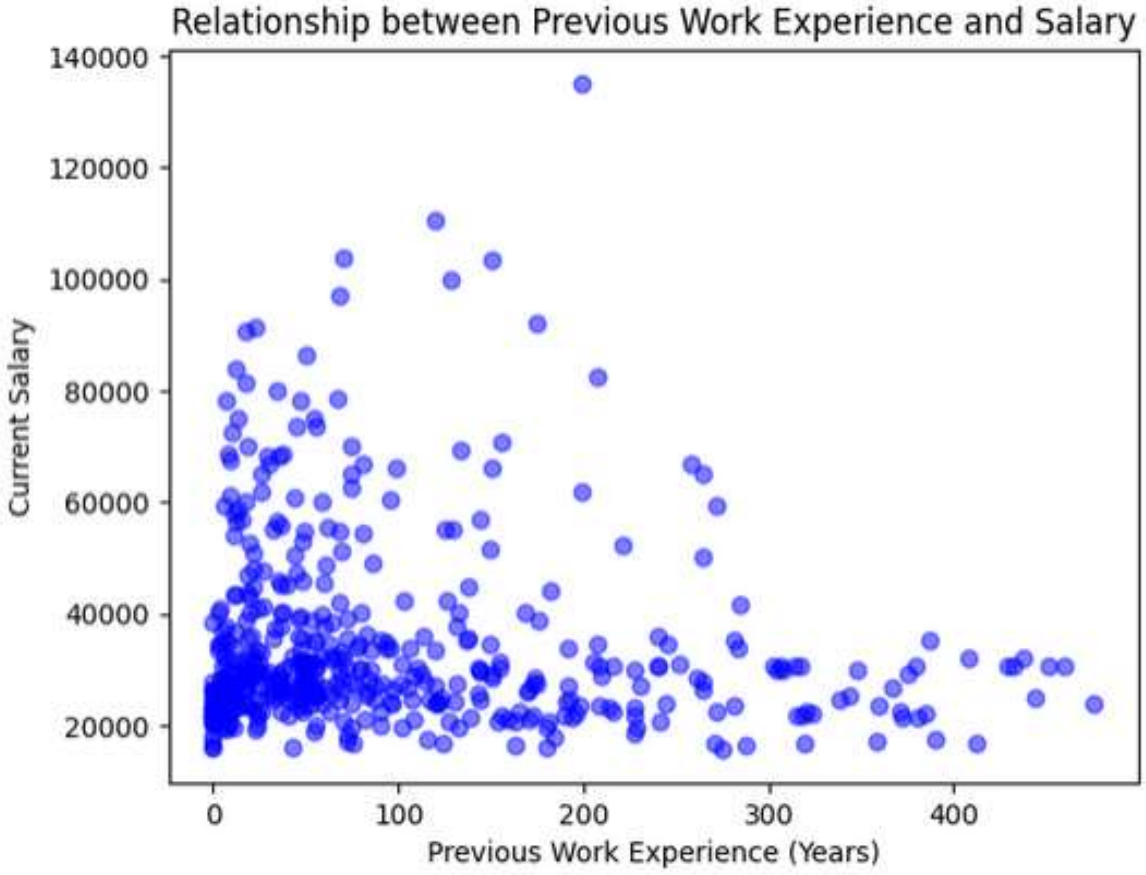
2) Generate a bar plot to compare the average salary of male and female employees using any library.

```
import matplotlib.pyplot as plt
import pandas as pd

df = pd.read_csv(r'C:\Users\dubey\Downloads\Employee data.csv')

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

plt.bar(gender_salary_avg.index, gender_salary_avg, color=['pink', 'red'])

plt.title('Average Salary Comparison by Gender')
plt.xlabel('Gender')
plt.ylabel('Average Salary')

plt.show()
```



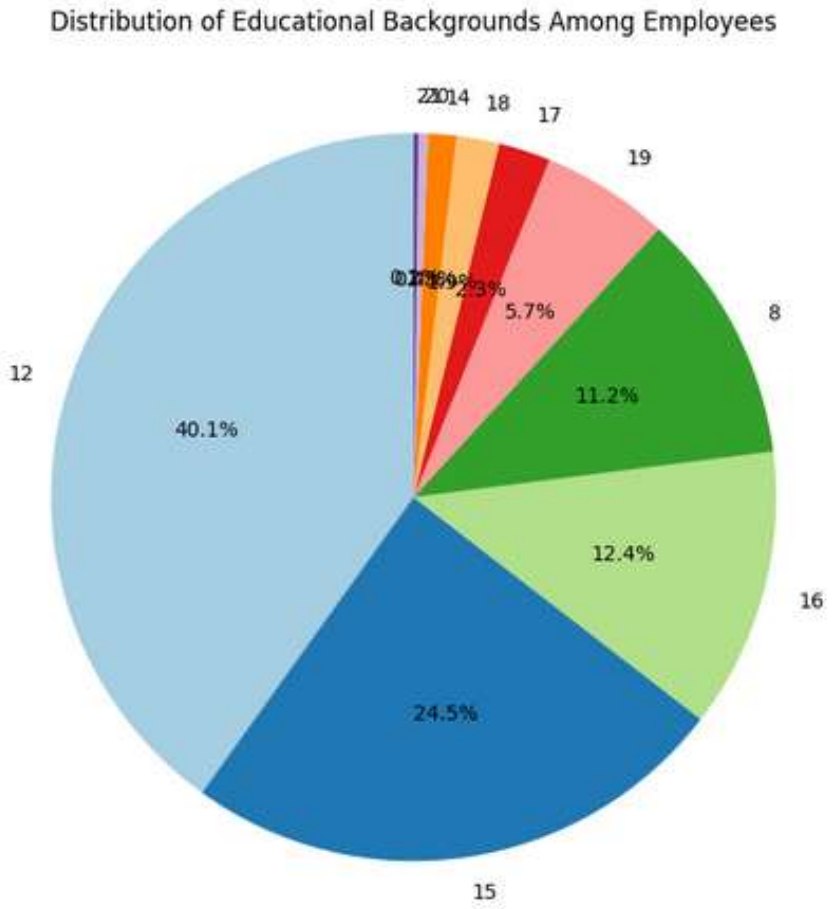
3) Create a scatter plot using any library to illustrate the relationship between previous work experience (prevexp) and the current salary of employees.

```
import matplotlib.pyplot as plt
import pandas as pd
df = pd.read_csv(r'C:\Users\dubey\Downloads\Employee data.csv')
plt.scatter(df['prevexp'], df['salary'], color='blue', alpha=0.5
plt.title('Relationship between Previous Work Experience and Salary')
plt.xlabel('Previous Work Experience (Years)')
plt.ylabel('Current Salary'
plt.show()
```



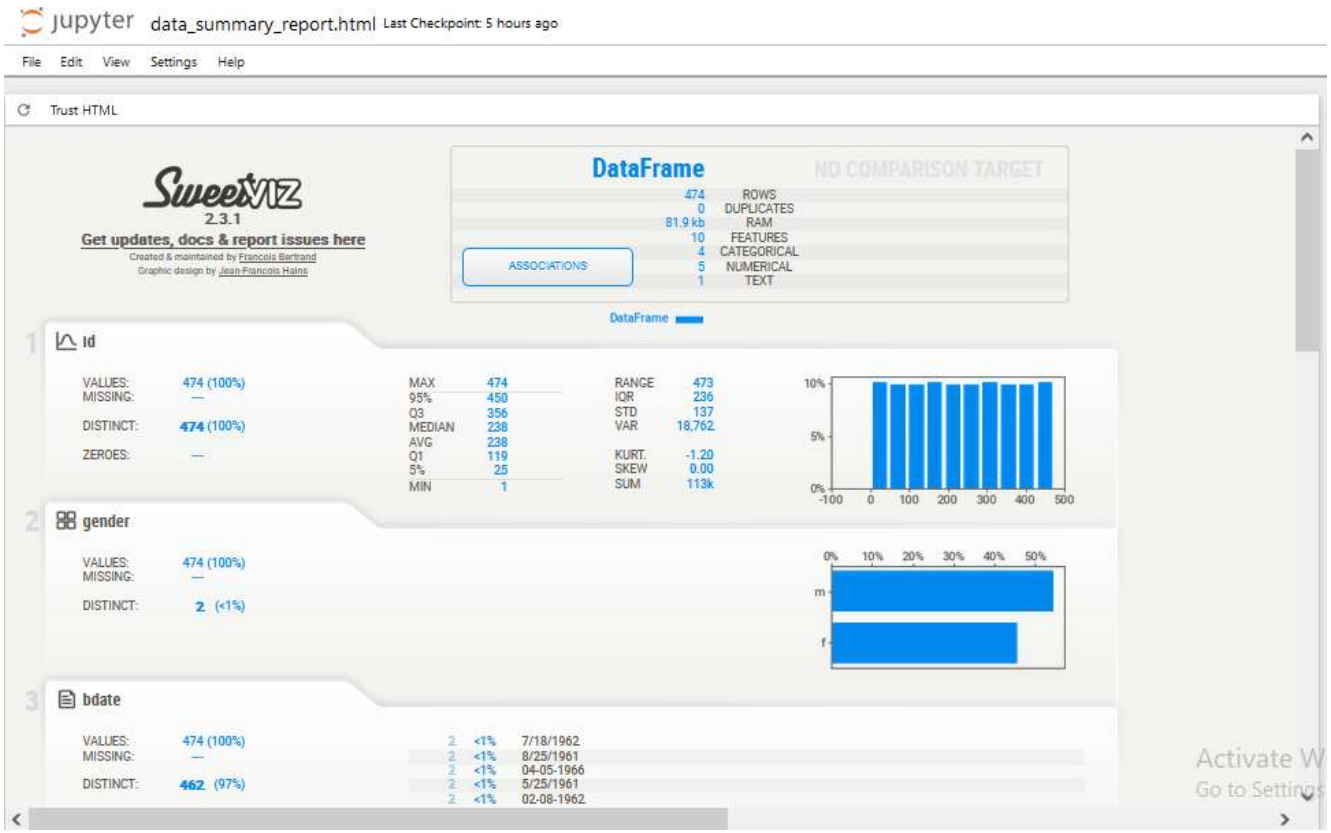
4)Generate a pie chart to visualize the distribution of educational backgrounds among employees.

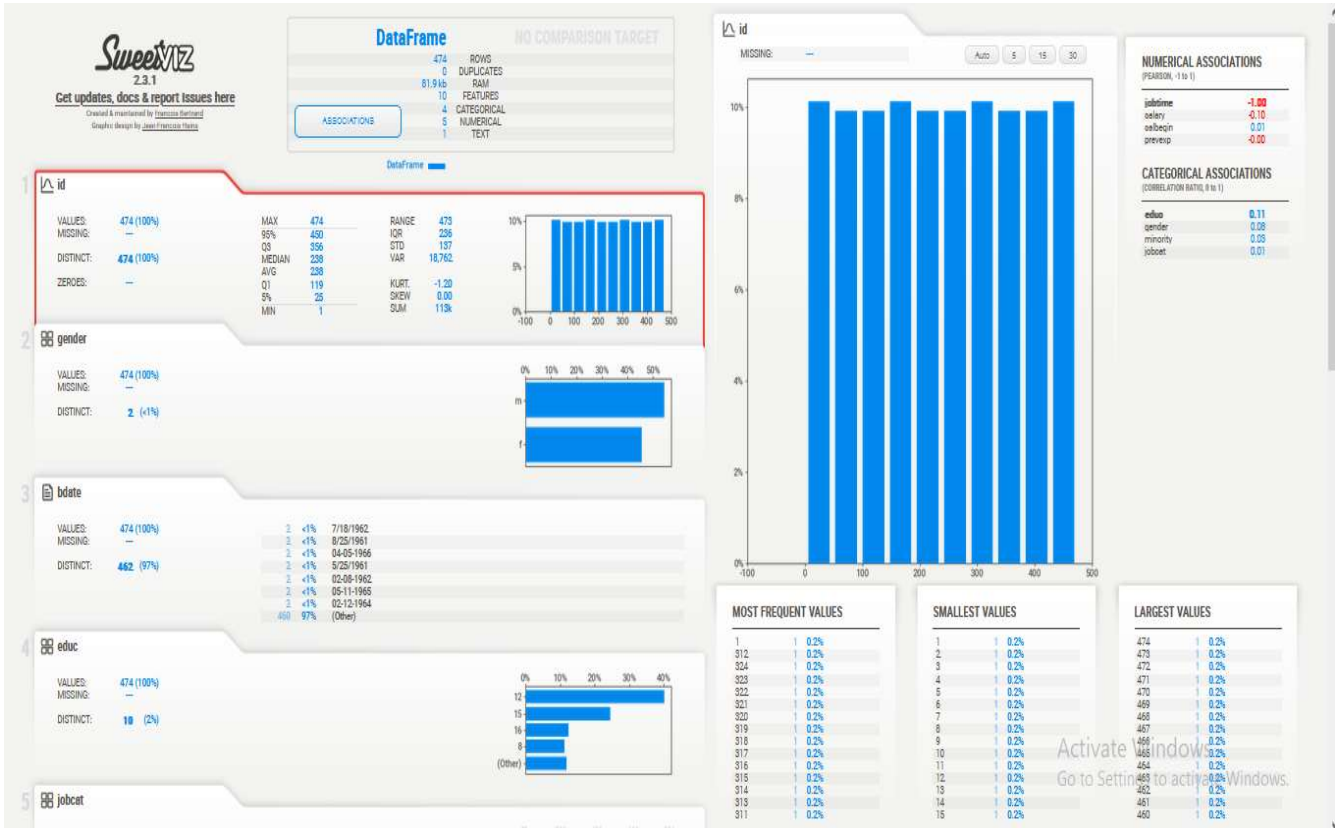
```
import matplotlib.pyplot as plt
import pandas as pd
df = pd.read_csv(r'C:\Users\dubey\Downloads\Employee data.csv')
education_distribution = df['educ'].value_counts()
plt.figure(figsize=(8, 8))
plt.pie(education_distribution,labels=education_distribution.index,
autopct='%1.1f%%', startangle=90, colors=plt.cm.Paired.colors)
plt.title('Distribution of Educational Backgrounds Among Employees')
plt.show()
```

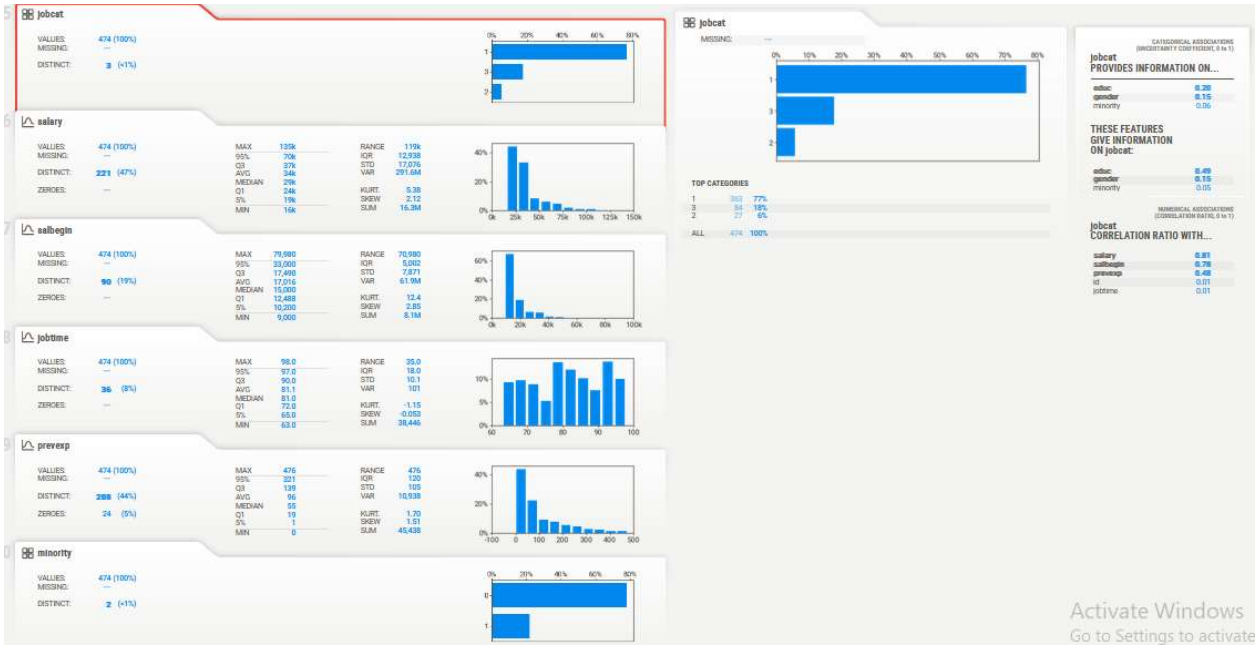


5) Use any AI library such as sweetviz or dtale to generate the summary of the data.

```
import sweetviz as sv

import pandas as pd

# Replace 'your_data.csv' with the actual file path

df = pd.read_csv(r'C:\Users\dubey\Downloads\Employee data.csv')


# Generate the summary report using sweetviz

report = sv.analyze(df)


# Save the report to an HTML file

report.show_html('data_summary_report.html')
```




SweetViz
2.3.1
Get updates, docs & report issues here
DataFrame
NO COMPARISON TARGET
474 ROWS
0 DUPLICATES
81.9 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%)
m
f
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: —
DISTINCT: 10 (2%)
12
15
16
8
(Other)
5 jobcat
id
MISSING: —
Auto
5
15
30
NUMERICAL ASSOCIATIONS
(PEARSON, -1 to 1)
jobtime -1.00
salary -0.10
salbegin 0.01
prevexp -0.00
CATEGORICAL ASSOCIATIONS
(CORRELATION RATIO, 0 to 1)
educ 0.11
gender 0.08
minority 0.03
jobcat 0.01
MOST FREQUENT VALUES
SMALLEST VALUES
LARGEST VALUES
Activate Windows
Go to Settings to activate Windows.

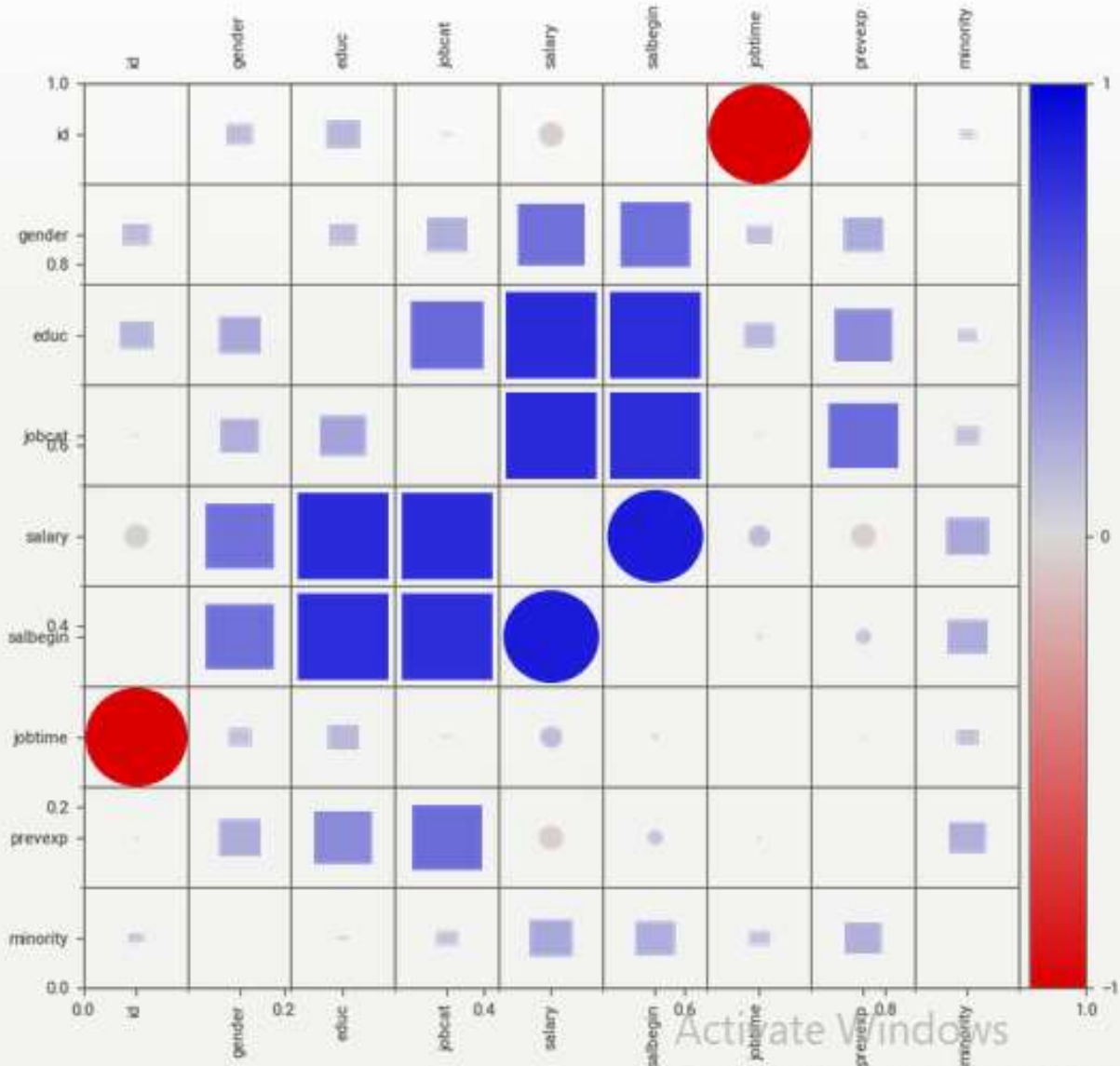

5 jobcat
VALUES: 474 (100%)
MISSING: —
DISTINCT: 3 (<1%)
6 salary
VALUES: 474 (100%)
MISSING: —
DISTINCT: 221 (47%)
ZEROES: —
MAX 135k
95% 70k
Q3 37k
AVG 34k
MEDIAN 29k
Q1 24k
5% 19k
MIN 16k
RANGE 119k
IQR 12,938
STD 17,076
VAR 291.6M
KURT. 5.38
SKEW 2.12
SUM 16.3M
7 salbegin
VALUES: 474 (100%)
MISSING: —
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
8 jobtime
VALUES: 474 (100%)
MISSING: —
DISTINCT: 36 (8%)
ZEROES: —
MAX 98.0
95% 87.0
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
9 prevexp
VALUES: 474 (100%)
MISSING: —
DISTINCT: 208 (44%)
ZEROES: 24 (5%)
MAX 476
95% 321
Q3 139
AVG 96
MEDIAN 55
Q1 19
5% 1
MIN 0
RANGE 476
IQR 120
STD 105
VAR 10,938
KURT. 1.70
SKEW 1.51
SUM 45,438
0 minority
VALUES: 474 (100%)
MISSING: —
DISTINCT: 2 (<1%)
jobcat
MISSING: —
TOP CATEGORIES
1 363 77%
3 84 18%
2 27 6%
ALL 474 100%
jobcat
PROVIDES INFORMATION ON...
educ 0.28
gender 0.15
minority 0.06
THESE FEATURES
GIVE INFORMATION
ON jobcat:
educ 0.49
gender 0.15
minority 0.05
jobcat
CORRELATION RATIO WITH...
salary 0.81
salbegin 0.78
prevexp 0.48
id 0.01
jobtime 0.01
Activate Windows
Go to Settings to activate

Associations

[Only including dataset "DataFrame"]

- **Squares** are categorical associations (uncertainty coefficient & correlation ratio) from 0 to 1. The uncertainty coefficient is **asymmetrical** (i.e. ROW LABEL values indicate how much they PROVIDE INFORMATION to each LABEL at the TOP).

- **Circles** are the symmetrical numerical correlations (Pearson's) from -1 to 1. The **trivial diagonal** is intentionally left blank for clarity.



used dtale to generate the summary of the data.

import dtale

Launch D-Tale web interface

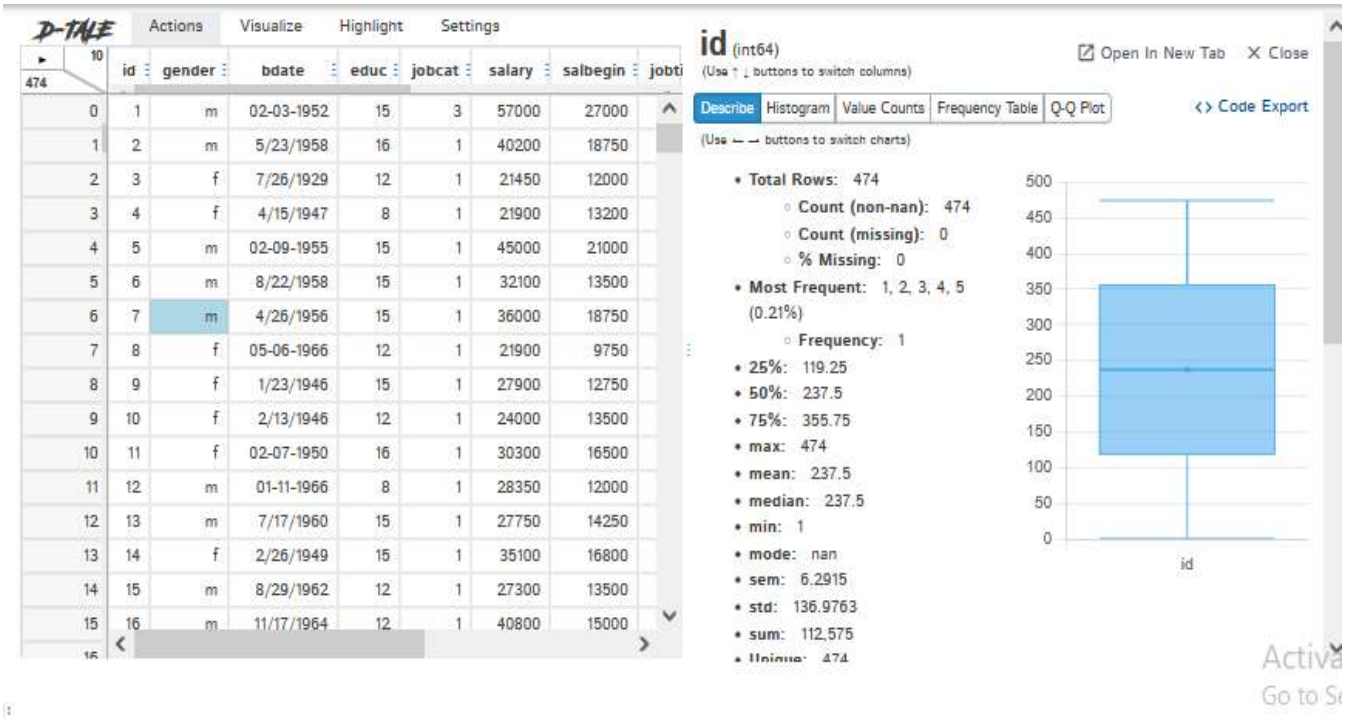
dtale.show(df)

D-TALE Actions Visualize Highlight Settings

10 / 474

	id	gender	bdate	educ	jobcat	salary	salbegin	jobtime	prevexp	minority
0	1	m	02-03-1952	15	3	57000	27000	98	144	0
1	2	m	5/23/1958	16	1	40200	18750	98	36	0
2	3	f	7/26/1929	12	1	21450	12000	98	381	0
3	4	f	4/15/1947	8	1	21900	13200	98	190	0
4	5	m	02-09-1955	15	1	45000	21000	98	138	0
5	6	m	8/22/1958	15	1	32100	13500	98	67	0
6	7	m	4/26/1956	15	1	36000	18750	98	114	0
7	8	f	05-06-1966	12	1	21900	9750	98	0	0
8	9	f	1/23/1946	15	1	27900	12750	98	115	0
9	10	f	2/13/1946	12	1	24000	13500	98	244	0
10	11	f	02-07-1950	16	1	30300	16500	98	143	0
11	12	m	01-11-1966	8	1	28350	12000	98	26	1
12	13	m	7/17/1960	15	1	27750	14250	98	34	1
13	14	f	2/26/1949	15	1	35100	16800	98	137	1
14	15	m	8/29/1962	12	1	27300	13500	97	66	0
15	16	m	11/17/1964	12	1	40800	15000	97	24	0
16	17	m	7/18/1962	15	1	46000	14250	97	48	0
17	18	m	3/20/1956	16	3	103750	27510	97	70	0
18	19	m	8/19/1962	12	1	42300	14250	97	103	0
19	20	f	1/23/1940	12	1	26250	11550	97	48	0
20	21	f	2/19/1963	16	1	38850	15000	97	17	0
21	22	m	9/24/1940	12	1	21750	12750	97	315	1

Describe (column analysis):



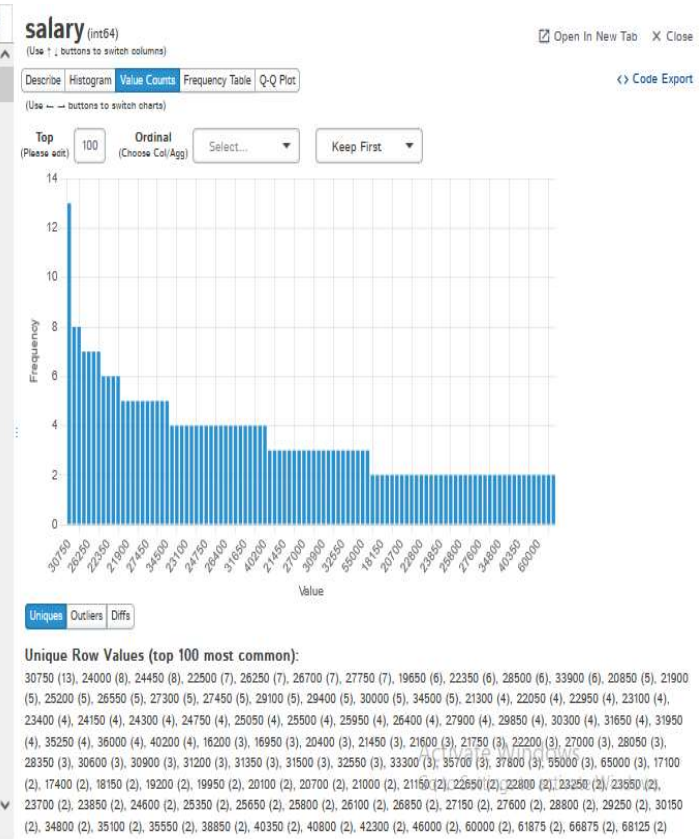
	id	gender	bdate	educ	jobcat	salary	salbegin	jobti
0	1	m	02-03-1952	15	3	57000	27000	
1	2	m	5/23/1958	16	1	40200	18750	
2	3	f	7/26/1929	12	1	21450	12000	
3	4	f	4/15/1947	8	1	21900	13200	
4	5	m	02-09-1955	15	1	45000	21000	
5	6	m	8/22/1958	15	1	32100	13500	
6	7	m	4/26/1956	15	1	36000	18750	
7	8	f	05-06-1966	12	1	21900	9750	
8	9	f	1/23/1946	15	1	27900	12750	
9	10	f	2/13/1946	12	1	24000	13500	
10	11	f	02-07-1950	16	1	30300	16500	
11	12	m	01-11-1966	8	1	28350	12000	
12	13	m	7/17/1960	15	1	27750	14250	
13	14	f	2/26/1949	15	1	35100	16800	
14	15	m	8/29/1962	12	1	27300	13500	
15	16	m	11/17/1964	12	1	40800	15000	

Histogram:

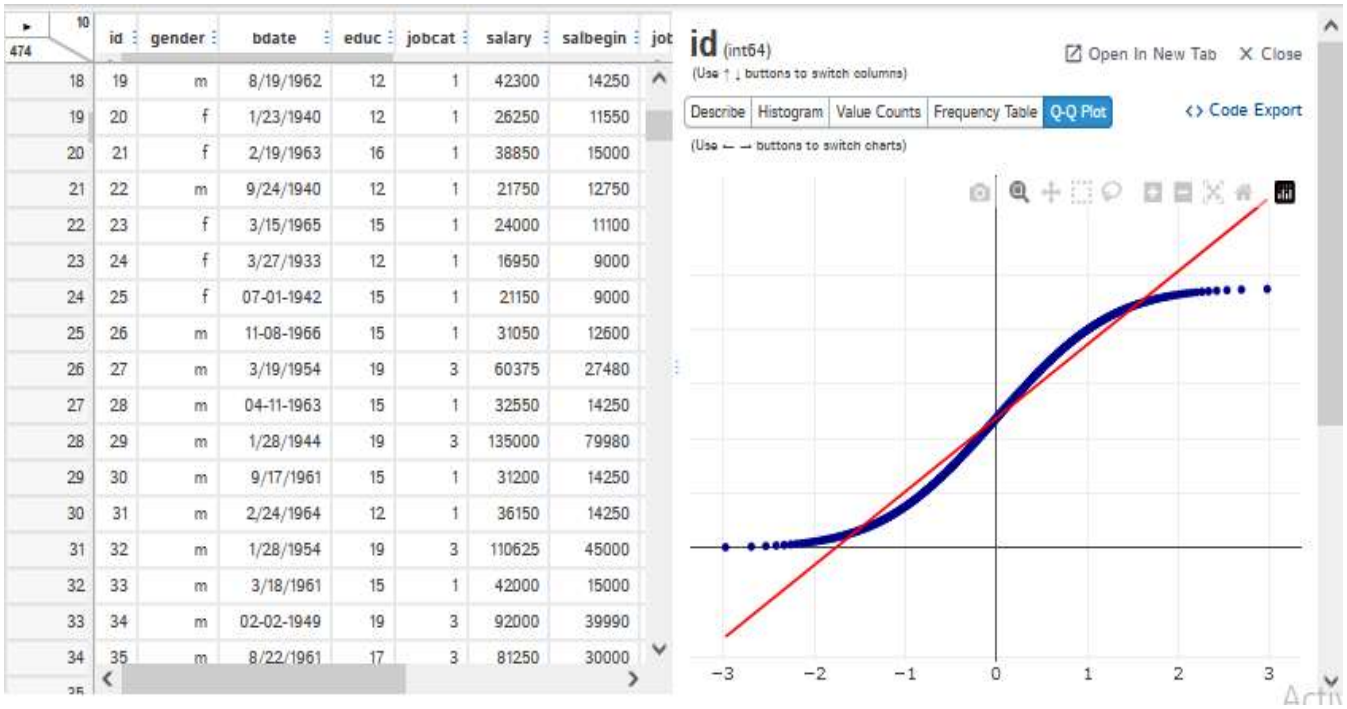
	id	gender	bdate	educ	jobcat	salary	salbegin	job
18	19	m	8/19/1962	12	1	42300	14250	
19	20	f	1/23/1940	12	1	26250	11550	
20	21	f	2/19/1963	16	1	38850	15000	
21	22	m	9/24/1940	12	1	21750	12750	
22	23	f	3/15/1965	15	1	24000	11100	
23	24	f	3/27/1933	12	1	16950	9000	
24	25	f	07-01-1942	15	1	21150	9000	
25	26	m	11-08-1966	15	1	31050	12600	
26	27	m	3/19/1954	19	3	60375	27480	
27	28	m	04-11-1963	15	1	32550	14250	
28	29	m	1/28/1944	19	3	135000	79980	
29	30	m	9/17/1961	15	1	31200	14250	
30	31	m	2/24/1964	12	1	36150	14250	
31	32	m	1/28/1954	19	3	110625	45000	
32	33	m	3/18/1961	15	1	42000	15000	
33	34	m	02-02-1949	19	3	92000	39990	
34	35	m	8/22/1961	17	3	81250	30000	

Value Count:

	id	gender	bdate	educ	jobcat	salary	salbegin	jobtime	prevexp	minority
0	1	m	02-03-1952	15	3	57000	27000	98	144	0
1	2	m	5/23/1958	16	1	40200	18750	98	36	0
2	3	f	7/26/1929	12	1	21450	12000	98	381	0
3	4	f	4/15/1947	8	1	21900	13200	98	190	0
4	5	m	02-09-1955	15	1	45000	21000	98	138	0
5	6	m	8/22/1958	15	1	32100	13500	98	67	0
6	7	m	4/26/1956	15	1	36000	18750	98	114	0
7	8	f	05-06-1966	12	1	21900	9750	98	0	0
8	9	f	1/23/1946	15	1	27900	12750	98	115	0
9	10	f	2/13/1946	12	1	24000	13500	98	244	0
10	11	f	02-07-1950	16	1	30300	16500	98	143	0
11	12	m	01-11-1966	8	1	28350	12000	98	26	1
12	13	m	7/17/1960	15	1	27750	14250	98	34	1
13	14	f	2/26/1949	15	1	35100	16800	98	137	1
14	15	m	8/29/1962	12	1	27300	13500	97	66	0
15	16	m	11/17/1964	12	1	40800	15000	97	24	0
16	17	m	7/18/1962	15	1	46000	14250	97	48	0
17	18	m	3/20/1956	16	3	103750	27510	97	70	0
18	19	m	8/19/1962	12	1	42300	14250	97	103	0
19	20	f	1/23/1940	12	1	26250	11550	97	48	0
20	21	f	2/19/1963	16	1	38850	15000	97	17	0
21	22	m	9/24/1940	12	1	21750	12750	97	315	1
22	23	f	3/15/1965	15	1	24000	11100	97	75	1
23	24	f	3/27/1933	12	1	16950	9000	97	124	1
24	25	f	07-01-1942	15	1	21150	9000	97	171	1
25	26	m	11-08-1966	15	1	31050	12600	96	14	0
26	27	m	3/19/1954	19	3	60375	27480	96	96	0
27	28	m	04-11-1963	15	1	32550	14250	96	43	0

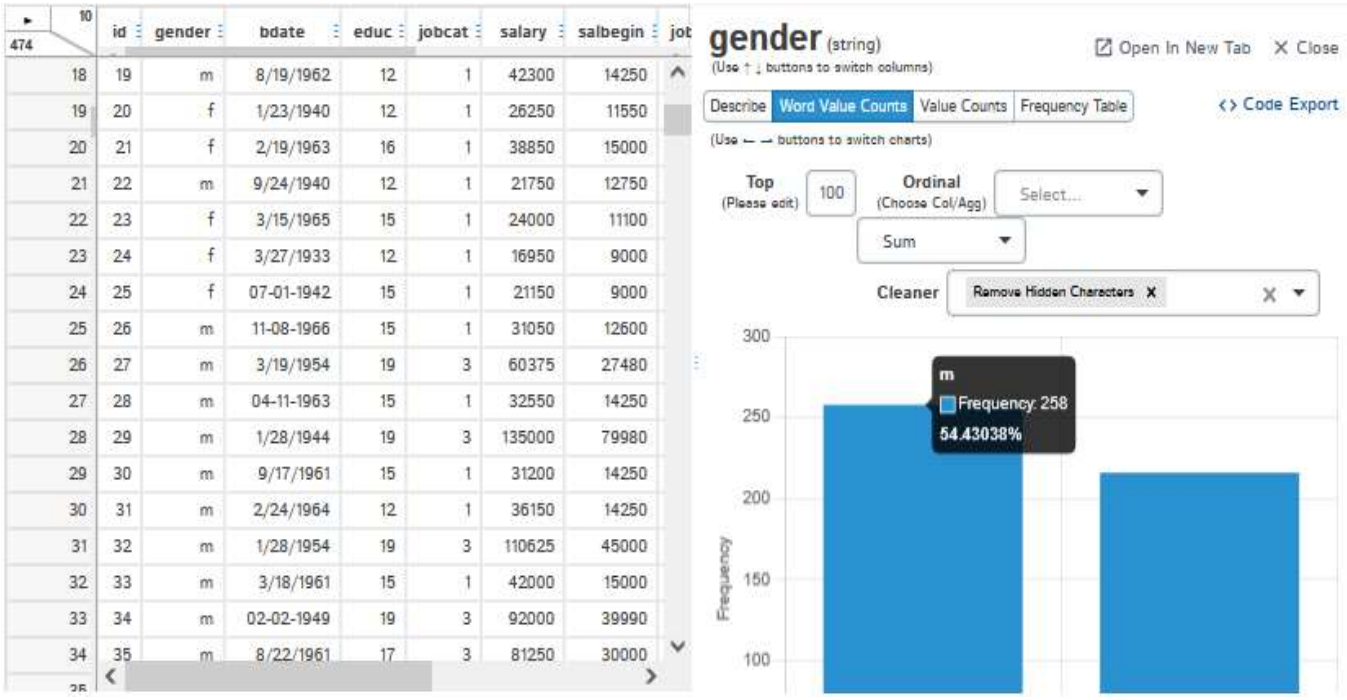


Q-Q plot:



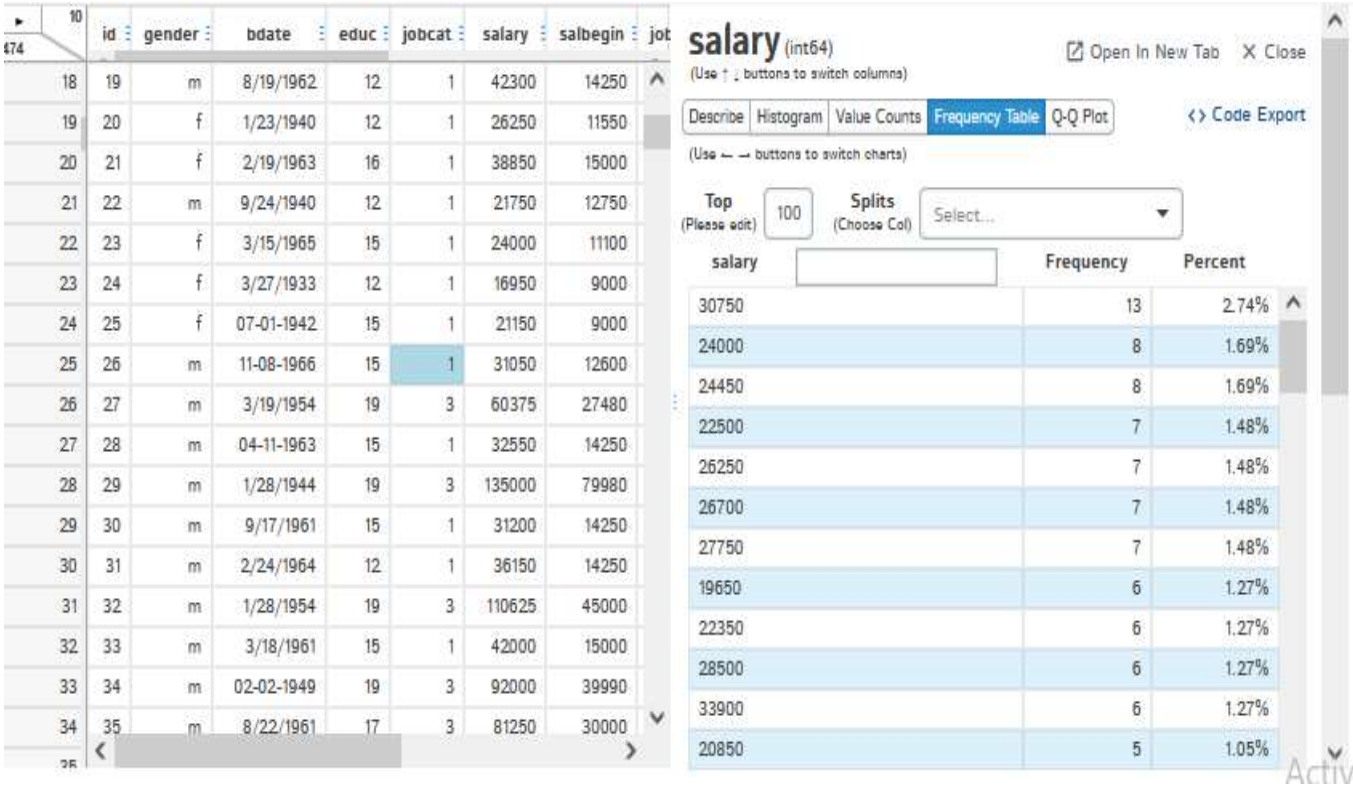
	id	gender	bdate	educ	jobcat	salary	salbegin	jot
18	19	m	8/19/1962	12	1	42300	14250	
19	20	f	1/23/1940	12	1	26250	11550	
20	21	f	2/19/1963	16	1	38850	15000	
21	22	m	9/24/1940	12	1	21750	12750	
22	23	f	3/15/1965	15	1	24000	11100	
23	24	f	3/27/1933	12	1	16950	9000	
24	25	f	07-01-1942	15	1	21150	9000	
25	26	m	11-08-1966	15	1	31050	12600	
26	27	m	3/19/1954	19	3	60375	27480	
27	28	m	04-11-1963	15	1	32550	14250	
28	29	m	1/28/1944	19	3	135000	79980	
29	30	m	9/17/1961	15	1	31200	14250	
30	31	m	2/24/1964	12	1	36150	14250	
31	32	m	1/28/1954	19	3	110625	45000	
32	33	m	3/18/1961	15	1	42000	15000	
33	34	m	02-02-1949	19	3	92000	39990	
34	35	m	8/22/1961	17	3	81250	30000	

Word value count:



	id	gender	bdate	educ	jobcat	salary	salbegin	jot
18	19	m	8/19/1962	12	1	42300	14250	
19	20	f	1/23/1940	12	1	26250	11550	
20	21	f	2/19/1963	16	1	38850	15000	
21	22	m	9/24/1940	12	1	21750	12750	
22	23	f	3/15/1965	15	1	24000	11100	
23	24	f	3/27/1933	12	1	16950	9000	
24	25	f	07-01-1942	15	1	21150	9000	
25	26	m	11-08-1966	15	1	31050	12600	
26	27	m	3/19/1954	19	3	60375	27480	
27	28	m	04-11-1963	15	1	32550	14250	
28	29	m	1/28/1944	19	3	135000	79980	
29	30	m	9/17/1961	15	1	31200	14250	
30	31	m	2/24/1964	12	1	36150	14250	
31	32	m	1/28/1954	19	3	110625	45000	
32	33	m	3/18/1961	15	1	42000	15000	
33	34	m	02-02-1949	19	3	92000	39990	
34	35	m	8/22/1961	17	3	81250	30000	

Frequency table:



	id	gender	bdate	educ	jobcat	salary	salbegin	jot
18	19	m	8/19/1962	12	1	42300	14250	
19	20	f	1/23/1940	12	1	26250	11550	
20	21	f	2/19/1963	16	1	38850	15000	
21	22	m	9/24/1940	12	1	21750	12750	
22	23	f	3/15/1965	15	1	24000	11100	
23	24	f	3/27/1933	12	1	16950	9000	
24	25	f	07-01-1942	15	1	21150	9000	
25	26	m	11-08-1966	15	1	31050	12600	
26	27	m	3/19/1954	19	3	60375	27480	
27	28	m	04-11-1963	15	1	32550	14250	
28	29	m	1/28/1944	19	3	135000	79980	
29	30	m	9/17/1961	15	1	31200	14250	
30	31	m	2/24/1964	12	1	36150	14250	
31	32	m	1/28/1954	19	3	110625	45000	
32	33	m	3/18/1961	15	1	42000	15000	
33	34	m	02-02-1949	19	3	92000	39990	
34	35	m	8/22/1961	17	3	81250	30000	

salary (int64)

(Use ↑ ↓ buttons to switch columns)

Open In New Tab Close

Describe | Histogram | Value Counts | Frequency Table | Q-Q Plot

Code Export

(Use ← → buttons to switch charts)

Top (Please edit) 100

Splits (Choose Col) Select...

salary	Frequency	Percent
30750	13	2.74%
24000	8	1.69%
24450	8	1.69%
22500	7	1.48%
26250	7	1.48%
26700	7	1.48%
27750	7	1.48%
19650	6	1.27%
22350	6	1.27%
28500	6	1.27%
33900	6	1.27%
20850	5	1.05%

D-TALE Actions Visualize Highlight Settings

474	id	gender	bdate	educ	jobcat	salary	salbegin	jobtime	prevexp	minority
448	449	m	01-02-1966	16	3	70000	21750	65	19	0
449	450	m	7/21/1954	19	3	55000	34980	65	129	0
450	451	m	7/19/1969	15	1	28500	14250	65	20	0
451	452	m	8/13/1953	12	1	28800	18000	65	210	0
452	453	m	08-07-1930	15	1	24450	15750	65	338	0
453	454	m	7/28/1965	19	3	90625	31250	65	18	0
454	455	m	1/17/1964	16	3	43650	19500	65	19	0
455	456	m	10/17/1959	19	3	75000	42510	65	54	0
456	457	m	5/27/1968	15	1	31650	14250	65	10	0
457	458	m	07-06-1965	19	3	61875	28740	65	26	0
458	459	f	02-10-1971	12	1	21750	11250	65	0	0
459	460	f	08-12-1969	12	1	22500	12750	65	24	0
460	461	f	11-08-1943	8	1	21600	13500	65	173	0
461	462	f	10/18/1963	16	3	34410	19500	65	79	0
462	463	f	10/15/1934	15	1	20700	14250	65	241	0
463	464	m	3/20/1962	19	3	47550	33000	64	27	0
464	465	m	7/20/1962	12	1	33900	16500	64	106	0
465	466	f	6/15/1948	12	1	23400	13500	64	198	0
466	467	f	8/18/1967	16	1	32850	19500	64	20	0
467	468	f	11/28/1965	16	3	55750	19980	64	36	0
468	469	f	06-01-1964	15	1	25200	13950	64	57	0
469	470	m	1/22/1964	12	1	26250	15750	64	69	1
470	471	m	08-03-1966	15	1	26400	15750	64	32	1
471	472	m	2/21/1966	15	1	39150	15750	63	46	0
472	473	f	11/25/1937	12	1	21450	12750	63	139	0

Conclusion:

In conclusion, the analysis of the employee dataset yielded crucial insights through a variety of visualizations. The histogram uncovered a broad spectrum of salary distributions, aiding in the identification of potential outliers and patterns. Gender-based salary differences were effectively highlighted using a bar plot, shedding light on areas for potential improvement in pay equity. The scatter plot provided a clear representation of the relationship between previous work experience and current salaries, offering valuable information for talent management and compensation strategies.

Furthermore, the distribution of educational backgrounds among employees was succinctly visualized using a pie chart, emphasizing the diverse academic backgrounds within the workforce. Finally, leveraging AI libraries such as sweetviz or dtale proved instrumental in generating concise data summaries, allowing for a quick grasp of key trends and statistical insights. These insights equip decision-makers with the necessary information to address issues related to compensation, diversity, and overall talent management effectively. The comprehensive analysis serves as a foundation for informed decision-making and strategic initiatives aimed at enhancing organizational performance and employee satisfaction.