Data preprocessing

SULTAN DANIYAR

Agenda:

- Loading data
- Data research
- Clearing data
- Data conversion

Loading data

- df = pd.read_csv('file_name') #read csv file
- df.head()#first 5(default) rows of df

Research

- column_names = df.columns
- column_data_types = df.dtypes
- print(column_names)
- print(column_data_types)

```
Index(['gender', 'age', 'Investment_Avenues', 'Mutual_Funds', 'Equity_Market',
       'Debentures', 'Government_Bonds', 'Fixed_Deposits', 'PPF', 'Gold',
       'Stock_Marktet', 'Factor', 'Objective', 'Purpose', 'Duration',
       'Invest_Monitor', 'Expect', 'Avenue',
       'What are your savings objectives?', 'Reason_Equity', 'Reason_Mutual',
       'Reason_Bonds', 'Reason_FD', 'Source'],
      dtype='object')
gender                                 object
age                                     int64
Investment_Avenues                     object
Mutual_Funds                            int64
Equity_Market                           int64
Debentures                              int64
Government_Bonds                        int64
Fixed_Deposits                          int64
PPF                                     int64
Gold                                    int64
Stock_Marktet                          object
Factor                                 object
Objective                              object
Purpose                                object
Duration                               object
Invest_Monitor                         object
Expect                                 object
Avenue                                 object
What are your savings objectives?      object
Reason_Equity                          object
Reason_Mutual                          object
Reason_Bonds                           object
Reason_FD                              object
Source                                 object
dtype: object
```

Research

- df_corr = df[['age','Mutual_Funds', 'Equity_Market']].corr()
- print(df_corr)

	age	Mutual_Funds	Equity_Market
age	1.000000	-0.123914	0.246840
Mutual_Funds	-0.123914	1.000000	0.332043
Equity_Market	0.246840	0.332043	1.000000

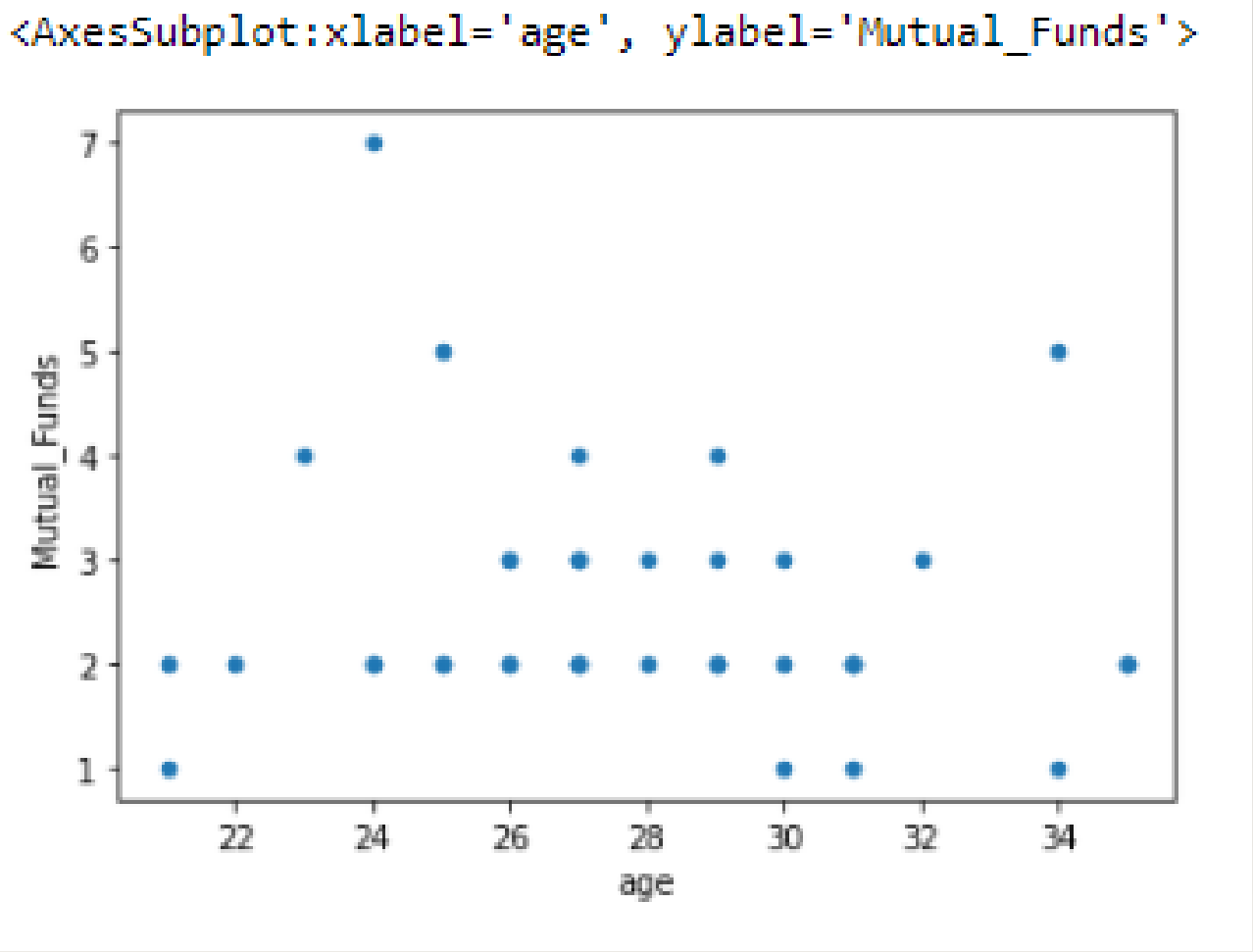
Research

- df_group = df.groupby('Mutual_Funds').max()['age']
- print(df_group)

```
Mutual_Funds
1     34
2     35
3     32
4     29
5     34
7     24
Name: AGE, dtype: int64
```

Visualizing

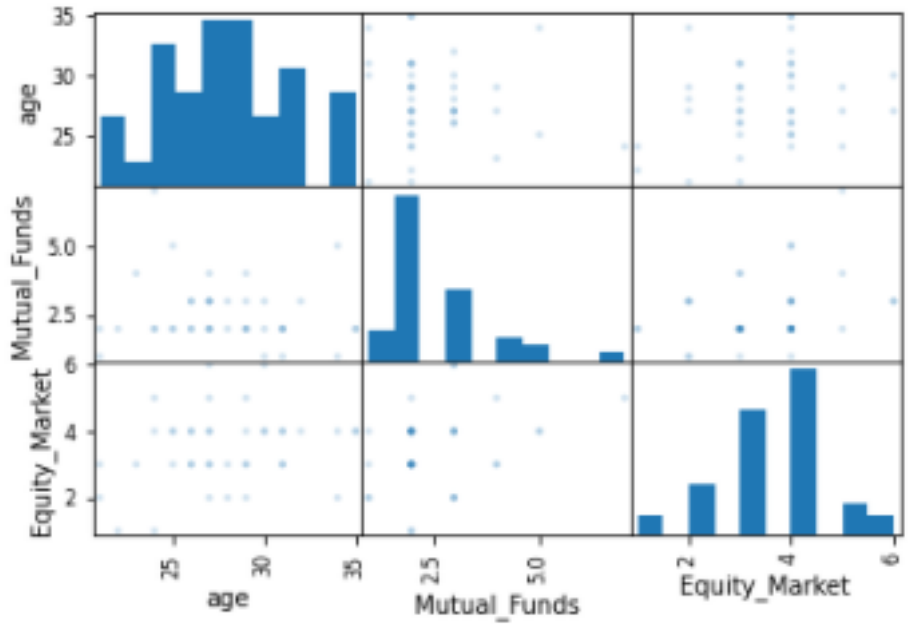
- df_sub = df[['age', 'Mutual_Funds']]
- df_sub.plot.scatter(x='age’,
- y='Mutual_Funds')



Matrix scatter

- from pandas.plotting import scatter_matrix
- df_sub1 = df[['age', 'Mutual_Funds’, 'Equity_Market']]
- scatter_matrix(df_sub1, alpha=0.2)

```
array([[<AxesSubplot:xlabel='age', ylabel='age'>,
        <AxesSubplot:xlabel='Mutual_Funds', ylabel='age'>,
        <AxesSubplot:xlabel='Equity_Market', ylabel='age'>],
       [<AxesSubplot:xlabel='age', ylabel='Mutual_Funds'>,
        <AxesSubplot:xlabel='Mutual_Funds', ylabel='Mutual_Funds'>,
        <AxesSubplot:xlabel='Equity_Market', ylabel='Mutual_Funds'>],
       [<AxesSubplot:xlabel='age', ylabel='Equity_Market'>,
        <AxesSubplot:xlabel='Mutual_Funds', ylabel='Equity_Market'>,
        <AxesSubplot:xlabel='Equity_Market', ylabel='Equity_Market'>]],
      dtype=object)
```



Null values

- df_missing = df.copy()
- df_missing.loc[0, 'age'] = np.nan
- print(df_missing[df_missing['age'].isnull()])

	gender	AGE	Investment_Avenues	Mutual_Funds	Equity_Market	Debentures	Government_Bonds	Fixed_Deposits	PPF	Gold	...	Invest_Monitor	Ex
0	Female	34	Yes	1	2	5	3	7	6	4	...	Monthly	20%
1	Female	23	Yes	4	3	2	1	5	6	7	...	Weekly	20%
2	Male	30	Yes	3	6	4	2	5	1	7	...	Daily	20%
3	Male	22	Yes	2	1	3	7	6	4	5	...	Daily	10%
4	Female	24	No	2	1	3	6	4	5	7	...	Daily	20%

Data filtering

- df.loc[df['AGE'] >= 15, ['AGE','Mutual_Funds']].head()

	AGE	Mutual_Funds
0	34	1
1	23	4
2	30	3
3	22	2
4	24	2

Columns rename

- df.rename(columns = {"age": "AGE"}, inplace = True)
- df.head()

	gender	AGE	Investment_Avenues	Mutual_Funds	Equity_Market	Debentures	Government_Bonds	Fixed_Deposits	PPF	Gold	...	Duration	Invest_Monitor
0	Female	34	Yes	1	2	5	3	7	6	4	...	1-3 years	Monthly
1	Female	23	Yes	4	3	2	1	5	6	7	...	More than 5 years	Weekly
2	Male	30	Yes	3	6	4	2	5	1	7	...	3-5 years	Daily
3	Male	22	Yes	2	1	3	7	6	4	5	...	Less than 1 year	Daily
4	Female	24	No	2	1	3	6	4	5	7	...	Less than 1 year	Daily

Creating new dataset

- new_dataset = df[['AGE', 'Mutual_Funds','Government_Bonds']]
- new_dataset.head()

	AGE	Mutual_Funds	Government_Bonds
0	34	1	3
1	23	4	1
2	30	3	2
3	22	2	7
4	24	2	6

Drop tables

- drop_df = df.drop(['Debentures', 'Government_Bonds'], axis=1)
- drop_df.head()

	gender	AGE	Investment_Avenues	Mutual_Funds	Equity_Market	Fixed_Deposits	PPF	Gold	Stock_Marktet	Factor	...	Duration	Invest_Monitor
0	Female	34	Yes	1	2	7	6	4	Yes	Returns	...	1-3 years	Monthly
1	Female	23	Yes	4	3	5	6	7	No	Locking Period	...	More than 5 years	Weekly
2	Male	30	Yes	3	6	5	1	7	Yes	Returns	...	3-5 years	Daily
3	Male	22	Yes	2	1	6	4	5	Yes	Returns	...	Less than 1 year	Daily
4	Female	24	No	2	1	4	5	7	No	Returns	...	Less than 1 year	Daily

THAT’S IT