# rk1-anderzzz

April 20, 2023

[30]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

[31]:
```
df = pd.read_csv('NIRS/toy_dataset.csv')
```

[32]:
```
df.shape
```

[32]:
```
(150000, 6)
```

[33]:
```
df.dtypes
```

[33]:
```
Number       int64
City        object
Gender      object
Age          int64
Income     float64
Illness     object
dtype: object
```

[34]:
```
df.nunique()
```

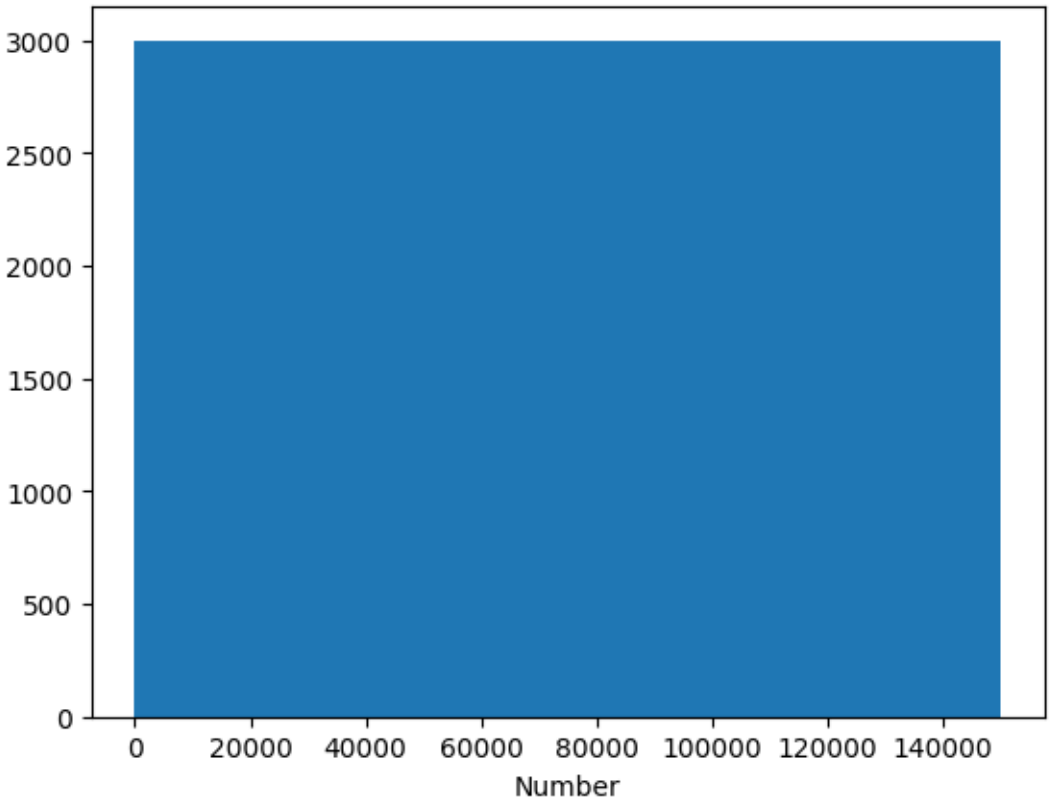
[34]:
```
Number     150000
City            8
Gender          2
Age            41
Income      71761
Illness         2
dtype: int64
```

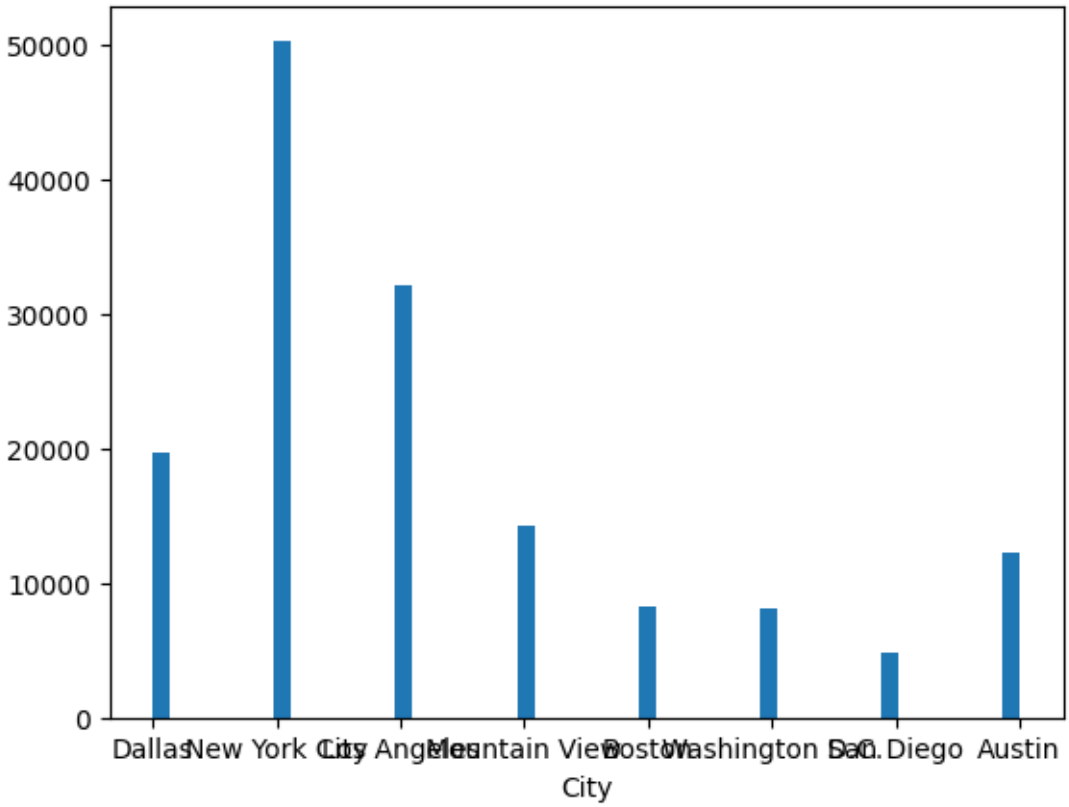
[35]:
```
df.isnull().sum()
```

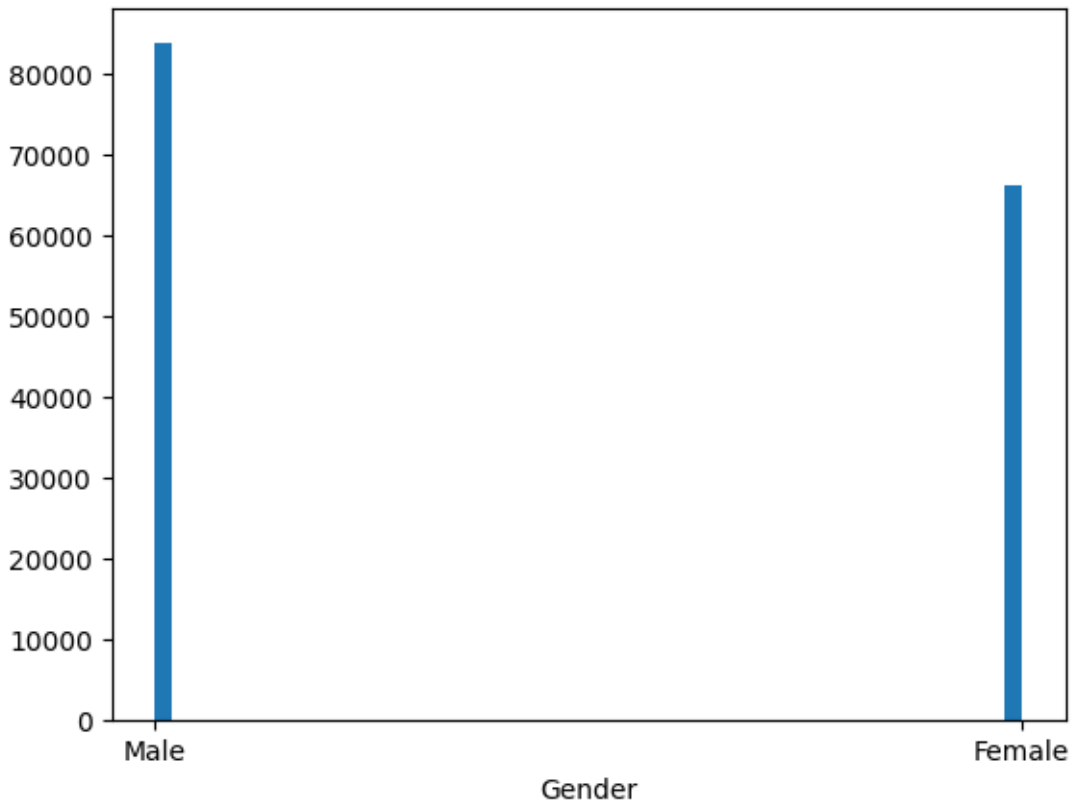
[35]:
```
Number     0
City       0
Gender     0
Age        0
```

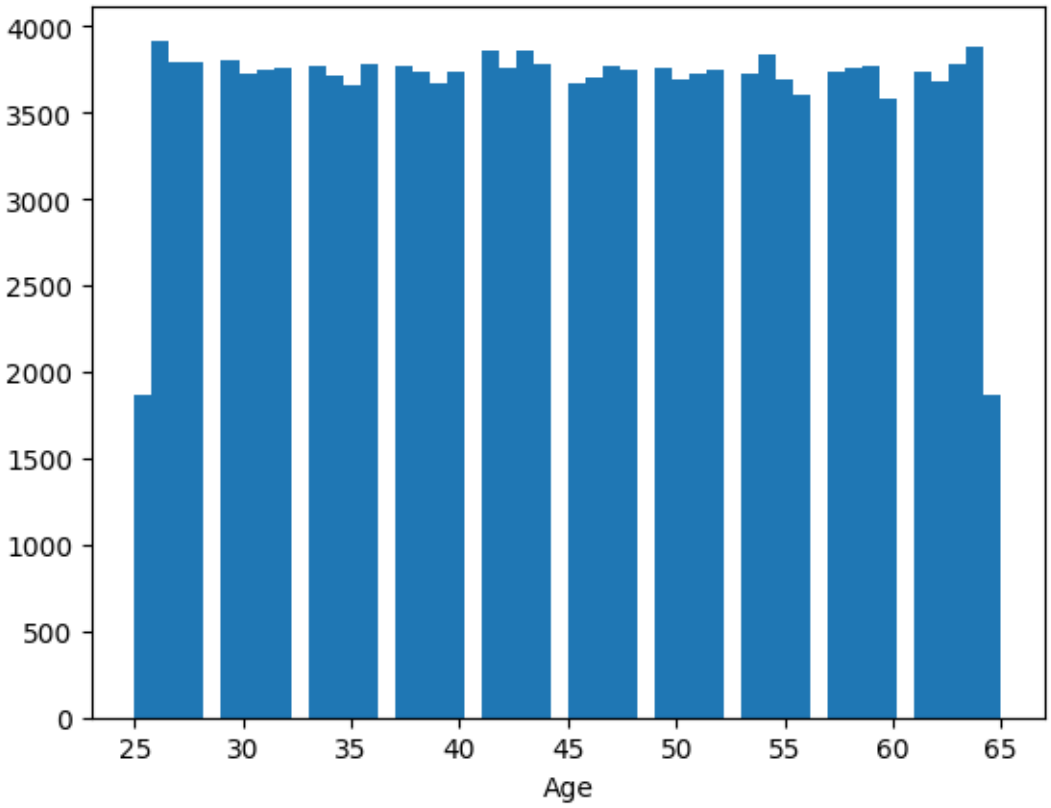
```
Income     0
Illness    0
dtype: int64
```

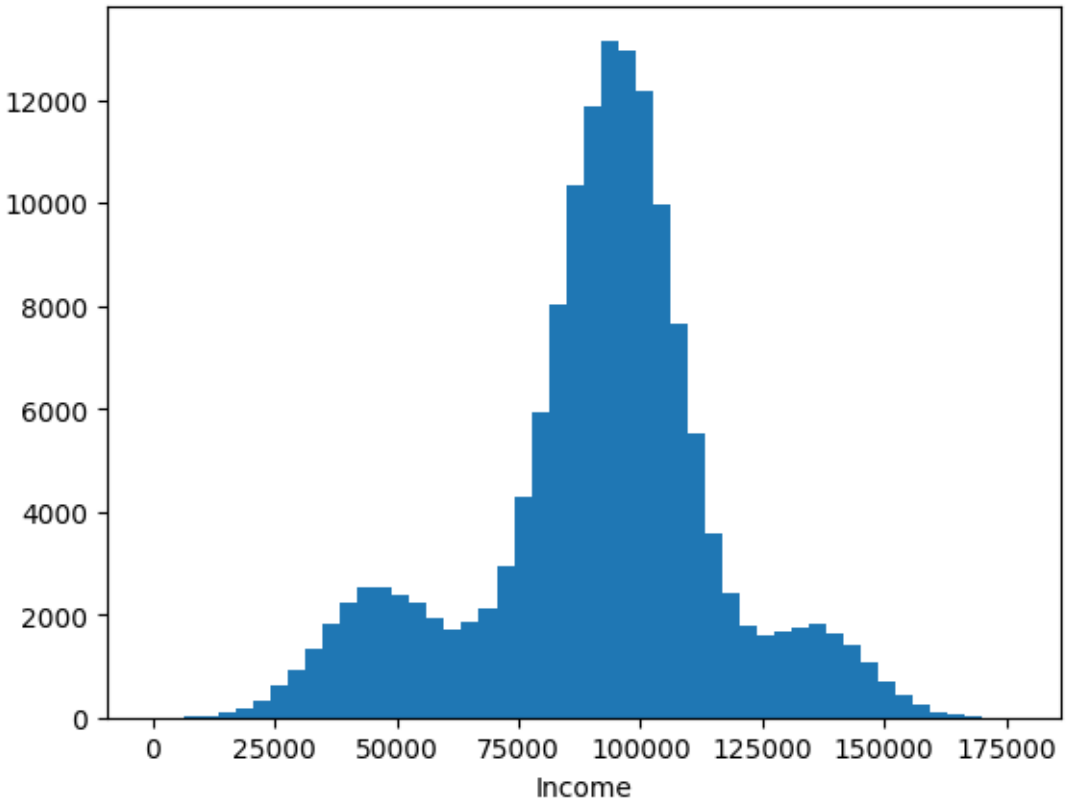
[36]:
```
for col in df:
    plt.hist(df[col], 50)
    plt.xlabel(col)
    plt.show()
```

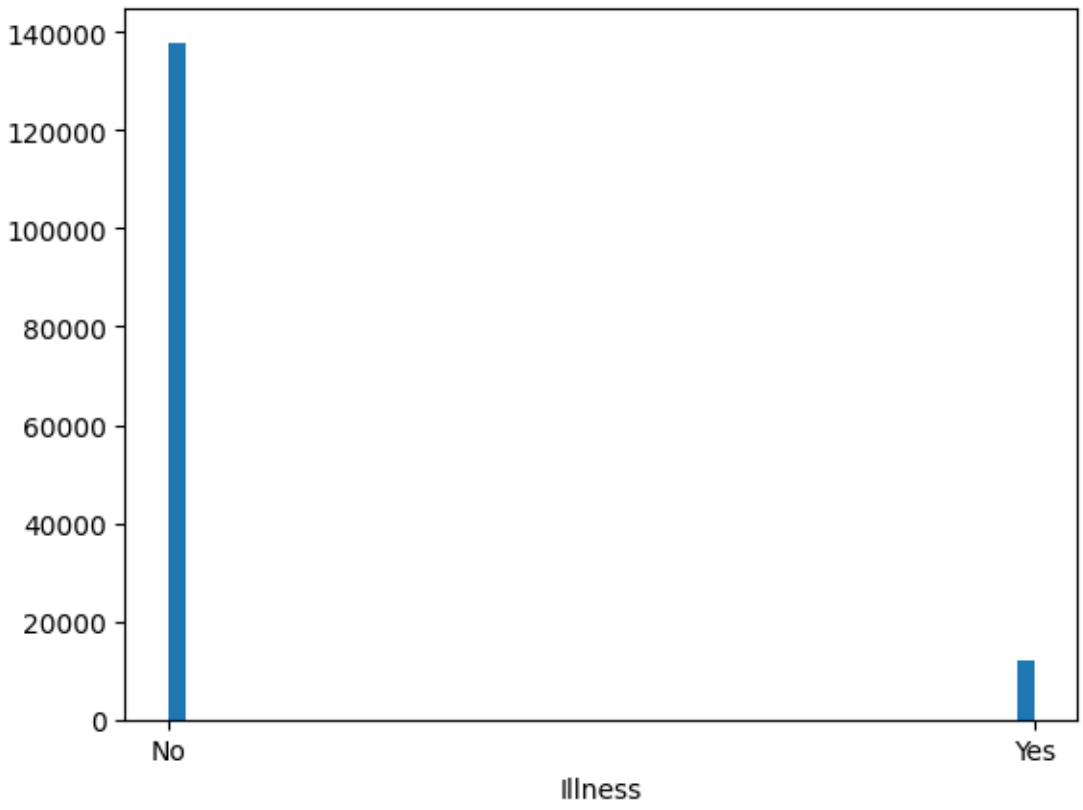



50000
40000
30000
20000
10000
0
Dallas
New York City
Los Angeles
Mountain View
Boston
Washington D.C.
San Diego
Austin
City


80000
70000
60000
50000
40000
30000
20000
10000
0
Male
Female
Gender


4000
3500
3000
2500
2000
1500
1000
500
0
25
30
35
40
45
50
55
60
65
Age


12000
10000
8000
6000
4000
2000
0
0
25000
50000
75000
100000
125000
150000
175000
Income




[37]:
```
df['Gender'].mask(df['Gender'] == 'Female', 0, inplace=True)
df['Gender'].mask(df['Gender'] == 'Male', 1, inplace=True)
df['Illness'].mask(df['Illness'] == 'Yes', 1, inplace=True)
df['Illness'].mask(df['Illness'] == 'No', 0, inplace=True)
df['Gender'] = pd.to_numeric(df['Gender'], errors='coerce')
df['Illness'] = pd.to_numeric(df['Illness'], errors='coerce')
df = df.drop(columns='Number')
df.dtypes
df.head()
```

[37]:

| | City | Gender | Age | Income | Illness |
|---|---|---|---|---|---|
| 0 | Dallas | 1 | 41 | 40367.0 | 0 |
| 1 | Dallas | 1 | 54 | 45084.0 | 0 |
| 2 | Dallas | 1 | 42 | 52483.0 | 0 |
| 3 | Dallas | 1 | 40 | 40941.0 | 0 |
| 4 | Dallas | 1 | 46 | 50289.0 | 0 |

[38]:
```
df.head()
```

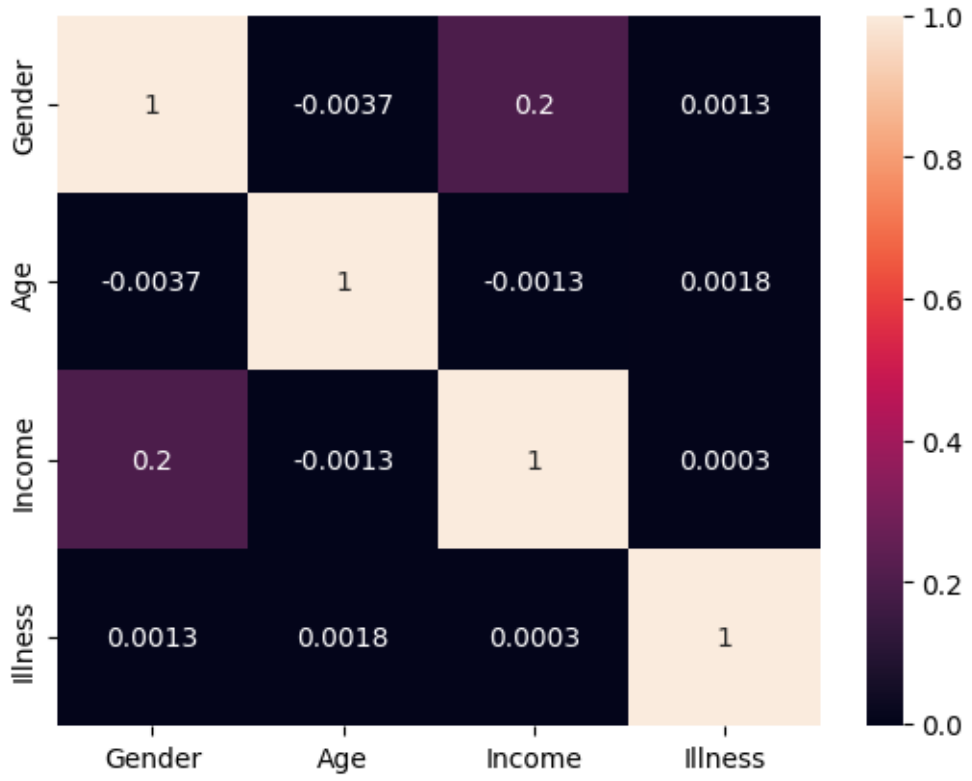
[38]:

| | City | Gender | Age | Income | Illness |
|---|---|---|---|---|---|
| 0 | Dallas | 1 | 41 | 40367.0 | 0 |

```
1  Dallas       1   54  45084.0        0
2  Dallas       1   42  52483.0        0
3  Dallas       1   40  40941.0        0
4  Dallas       1   46  50289.0        0
```

[39]:
```
df = df.drop(columns='City')
sns.heatmap(df.corr(), annot = True)
```

[39]:
```
<AxesSubplot: >
```



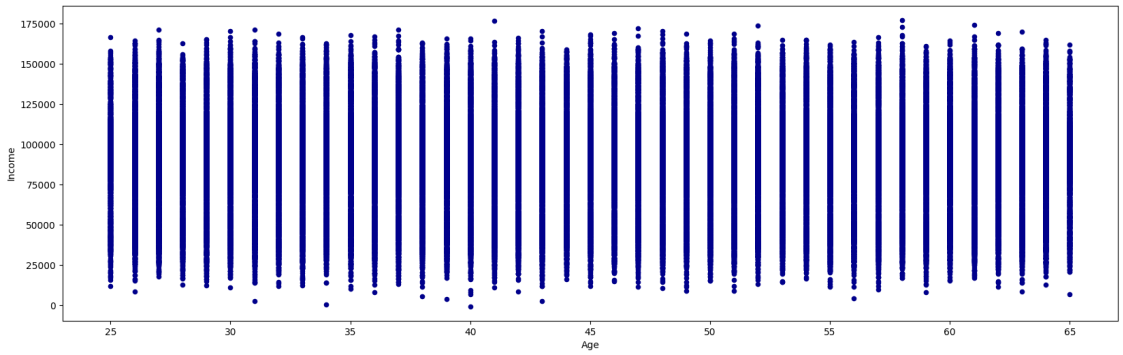

[50]:
```
df.index = map(int, Income.index)
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[50], line 1
----> 1 df.index = map(int, Income.index)

NameError: name 'Income' is not defined
```

[56]:
```
df.plot(kind='scatter', x = 'Age', y='Income', figsize=(20, 6),␣
↪color='darkblue')
```

```
plt.xlabel('Age')
plt.ylabel('Income')
plt.show()
```




[44]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150000 entries, 0 to 149999
Data columns (total 4 columns):
 #   Column   Non-Null Count   Dtype
---  ------   --------------   -----
 0   Gender   150000 non-null  int64
 1   Age      150000 non-null  int64
 2   Income   150000 non-null  float64
 3   Illness  150000 non-null  int64
dtypes: float64(1), int64(3)
memory usage: 4.6 MB
```

[45]:
```
df.head()
```

[45]:

| | Gender | Age | Income | Illness |
|---|---|---|---|---|
| 0 | 1 | 41 | 40367.0 | 0 |
| 1 | 1 | 54 | 45084.0 | 0 |
| 2 | 1 | 42 | 52483.0 | 0 |
| 3 | 1 | 40 | 40941.0 | 0 |
| 4 | 1 | 46 | 50289.0 | 0 |

[ ]:
```

```