

Overview Dataset

"Data Sensus Penghasilan". Data ini diambil dari database biro Sensus 1994 oleh Ronny Kohavi dan Barry Becker (Data Mining and Visualization, Silicon Graphics).

Data set ini berisikan beberapa kategori sebagai berikut

- age
- sex
- race
- education
- marital-status
- workclass
- occupation
- relationship
- native-country
- hours-per-week
- capital-loss
- capital-gain
- fnlwgt

Exploratory Data Analysis (EDA)

Pada praktikum ini, Anda akan melihat beberapa contoh Exploratory Data Analysis (EDA) pada data set sensus penghasilan.

- [Pengambilan Data](#)
- [Univariate analysis](#)
- [Bivariate analysis](#)

Instruksi Tugas Praktikum FIKTI dan FTI

1. Buat dan jelaskan univariate analysis untuk kategori hours-per-week dan income
2. Buat dan jelaskan Bivariate analysis untuk hubungan antara hours-per-week dan income

1. Pengambilan Data:

1.1 Import packages

```
In [10]: import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
import warnings
warnings.filterwarnings("ignore")
from scipy.stats import ttest_ind, ttest_rel
from scipy import stats
```

1.2 Import data

```
In [11]: data = pd.read_csv("https://gitlab.com/andreass.bayu/file-directory/-/raw/main/adult.csv")
data.head(10)
```

Out[11]:

	age	workclass	fnlwgt	education	educational-num	marital-status	occupation	relationship	race	gender	capital-gain	capital-loss
0	25	Private	226802	11th	7	Never-married	Machine-op-inspct	Own-child	Black	Male	0	0
1	38	Private	89814	HS-grad	9	Married-civ-spouse	Farming-fishing	Husband	White	Male	0	0
2	28	Local-gov	336951	Assoc-acdm	12	Married-civ-spouse	Protective-serv	Husband	White	Male	0	0
3	44	Private	160323	Some-college	10	Married-civ-spouse	Machine-op-inspct	Husband	Black	Male	7688	0
4	18	?	103497	Some-college	10	Never-married	?	Own-child	White	Female	0	0
5	34	Private	198693	10th	6	Never-married	Other-service	Not-in-family	White	Male	0	0
6	29	?	227026	HS-grad	9	Never-married	?	Unmarried	Black	Male	0	0
7	63	Self-emp-not-inc	104626	Prof-school	15	Married-civ-spouse	Prof-specialty	Husband	White	Male	3103	0
8	24	Private	369667	Some-college	10	Never-married	Other-service	Unmarried	White	Female	0	0
9	55	Private	104996	7th-8th	4	Married-civ-spouse	Craft-repair	Husband	White	Male	0	0

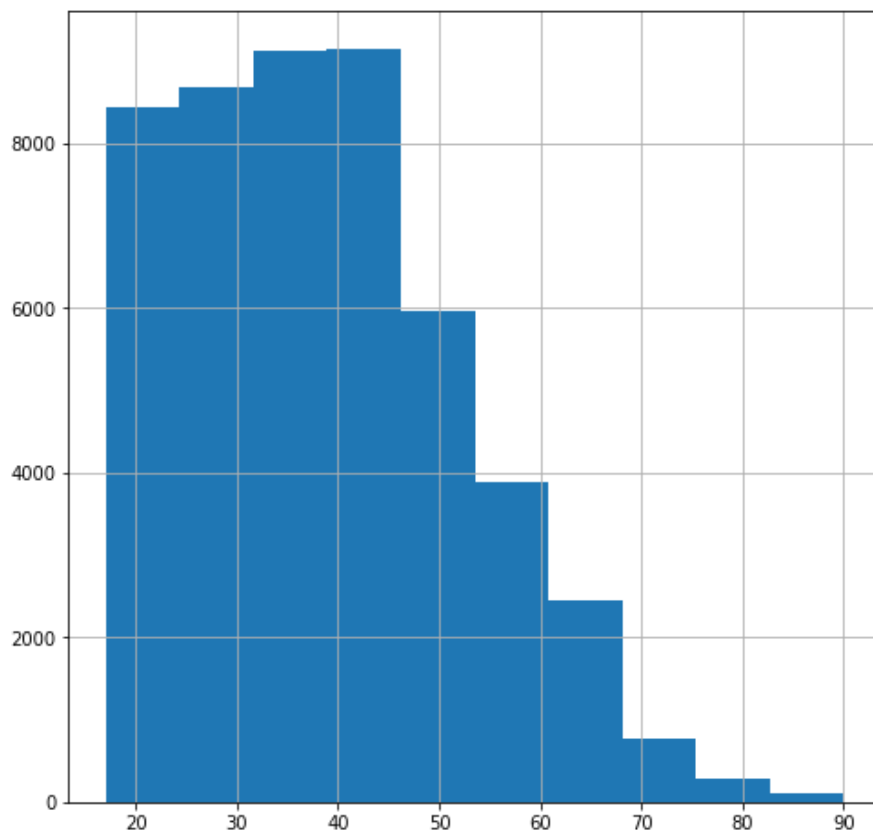
2. EDA

2.1.Univariate analysis

2.1.1 Age

i. Distribusi

```
In [5]: data['age'].hist(figsize=(8,8))  
plt.show()
```



```
In [11]: data[data["age"]>70].shape
```

```
Out[11]: (868, 15)
```

ii. Deskripsi

Histogram di atas menjelaskan bahwa :

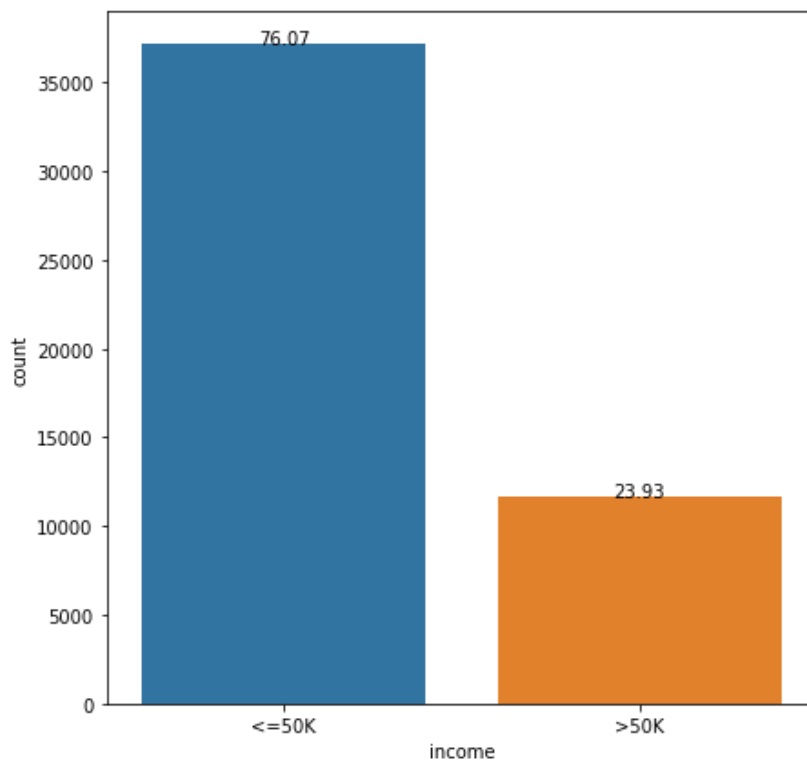
- atribut "usia" tidak simetris.
- miring ke kanan (Tapi ini baik karena orang dewasa yang lebih muda mendapatkan upah bukan yang lebih tua)
- Usia minimum dan maksimum orang adalah 17 dan 90 tahun.
- Dataset ini memiliki lebih sedikit pengamatan (868) tentang usia orang setelah usia tertentu, yaitu 70 tahun.

2.1.2 Income

i. Distribusi

```
In [12]: plt.figure(figsize=(7,7))
total = float(len(data) )

ax = sns.countplot(x="income", data=data)
for p in ax.patches:
    height = p.get_height()
    ax.text(p.get_x()+p.get_width()/2.,
            height + 3,
            '{:1.2f}'.format((height/total)*100),
            ha="center")
plt.show()
```



ii. Deskripsi

Distribusi di atas menjelaskan bahwa :

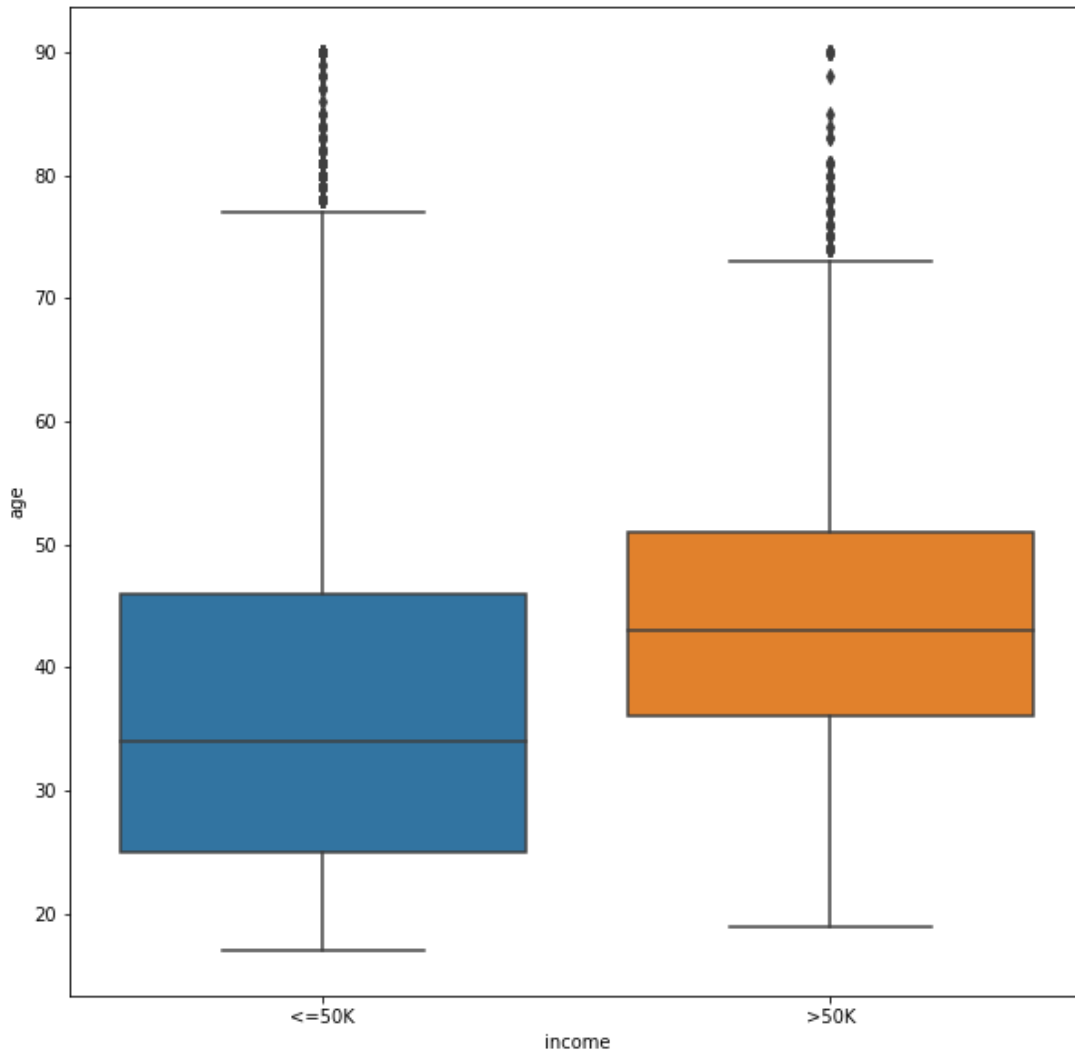
- Dataset ini tidak seimbang, yaitu 23,93% di antaranya termasuk dalam kelompok pendapatan 1 (yang berpenghasilan lebih dari \$50K) dan 76% termasuk dalam kelompok pendapatan 0 (yang berpenghasilan kurang dari \$50K).

2.2.Bivariate analysis

2.2.1 Age - income (Hubungan)

i. Boxplot

```
In [13]: fig = plt.figure(figsize=(10,10))
sns.boxplot(x="income", y="age", data=data)
plt.show()
```



```
In [14]: data[['income', 'age']].groupby(['income'], as_index=False).mean().sort_values(by='age', a
```

```
Out[14]:
```

	income	age
1	>50K	44.275178
0	<=50K	36.872184

ii. Deskripsi

Bivariate boxplot di atas menjelaskan bahwa :

- Rata-rata "usia" untuk kelompok Penghasilan (<= \$50K) adalah 36,8 tahun. Dan untuk kelompok Penghasilan (> \$50K) adalah 44,2 tahun

- Kelompok pendapatan($\leq 50K$) memiliki median "usia"(34 tahun) lebih rendah daripada kelompok Pendapatan($>50K$) yang memiliki median "usia"(42 tahun).

iii. Uji Hipotesis (untuk menguji hubungan antara Age & Income)

Two sampled T-test :-Uji Independent Samples t Test atau 2-sample t-test membandingkan rata-rata dua kelompok independen untuk menentukan apakah ada bukti statistik bahwa rata-rata populasi terkait berbeda secara signifikan.

Contoh : apakah ada hubungan antara usia dan pendapatan?

Tentukan null dan alternate hypothesis.

Secara umum, null hypothesis akan menyatakan bahwa kedua populasi yang diuji tidak memiliki perbedaan yang signifikan secara statistik.
alternate hypothesis akan menyatakan bahwa ada perbedaan.

Dalam contoh ini kita dapat mengatakan bahwa:

- Null Hypothesis :- tidak ada perbedaan Rerata usia kelompok pendapatan $>50k$ dan kelompok pendapatan $\leq 50k$.
- Alternate Hypothesis :- ada perbedaan Rerata usia kelompok pendapatan $>50k$ dan kelompok pendapatan $\leq 50k$.

```
In [33]: import random

data = data[(np.abs(stats.zscore(data["age"]))) < 3]]

income_1 = data[data['income']=='<=50K']['age']
income_0 = data[data['income']=='>50K']['age']

income_0 = income_0.values.tolist()
income_0 = random.sample(income_0, 100)
income_1 = income_1.values.tolist()
income_1 = random.sample(income_1, 100)
```

```
In [39]: from scipy.stats import ttest_ind
ttest,pval = ttest_ind(income_1,income_0,equal_var = False)
print("ttest",ttest)
print('p value',pval)

if pval < 0.05:
    print("null hypothesis ditolak")
else:
    print("null hypothesis diterima")

ttest -4.522346980129441
p value 1.1014567838186016e-05
null hypothesis ditolak
```

iv. Kesimpulan akhir

Menggunakan analisis statistik,

Dapat disimpulkan bahwa terdapat perbedaan yang signifikan pada rerata usia kelompok berpenghasilan >50K dan kelompok berpenghasilan <=50K.

Hal Ini membuktikan bahwa usia memiliki peranan dalam membedakan kelompok pendapatan.

Tugas

1. Univariate analysis

1.1 hours-per-week

i. Distribusi

```
In [1]: # Buat diagram hours-per-week
data['hours-per-week'].hist(figsize=(8,8))
plt.show
```

ii. Deskripsi

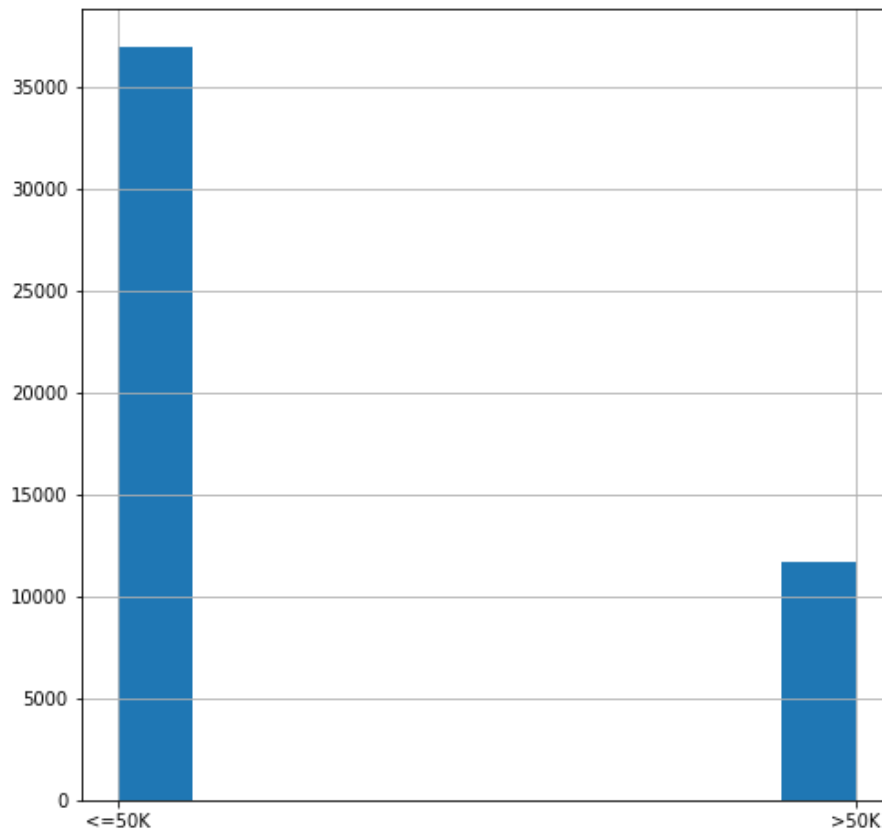
```
In [2]: # Jelaskan hasil diagram tersebut

# / *histogram diatas dapat menjelaskan bahwa: atribut "hours per week" termasuk simetris s
# dihasilkan diperoleh sebanyak 40 jam) Jam kerja minimum dan maksimum adalah 1 dan 99 jam
```

1.2 Income

i. Distribusi

```
In [23]: # Buat diagram income  
data['income'].hist(figsize=(8,8))  
plt.show()
```



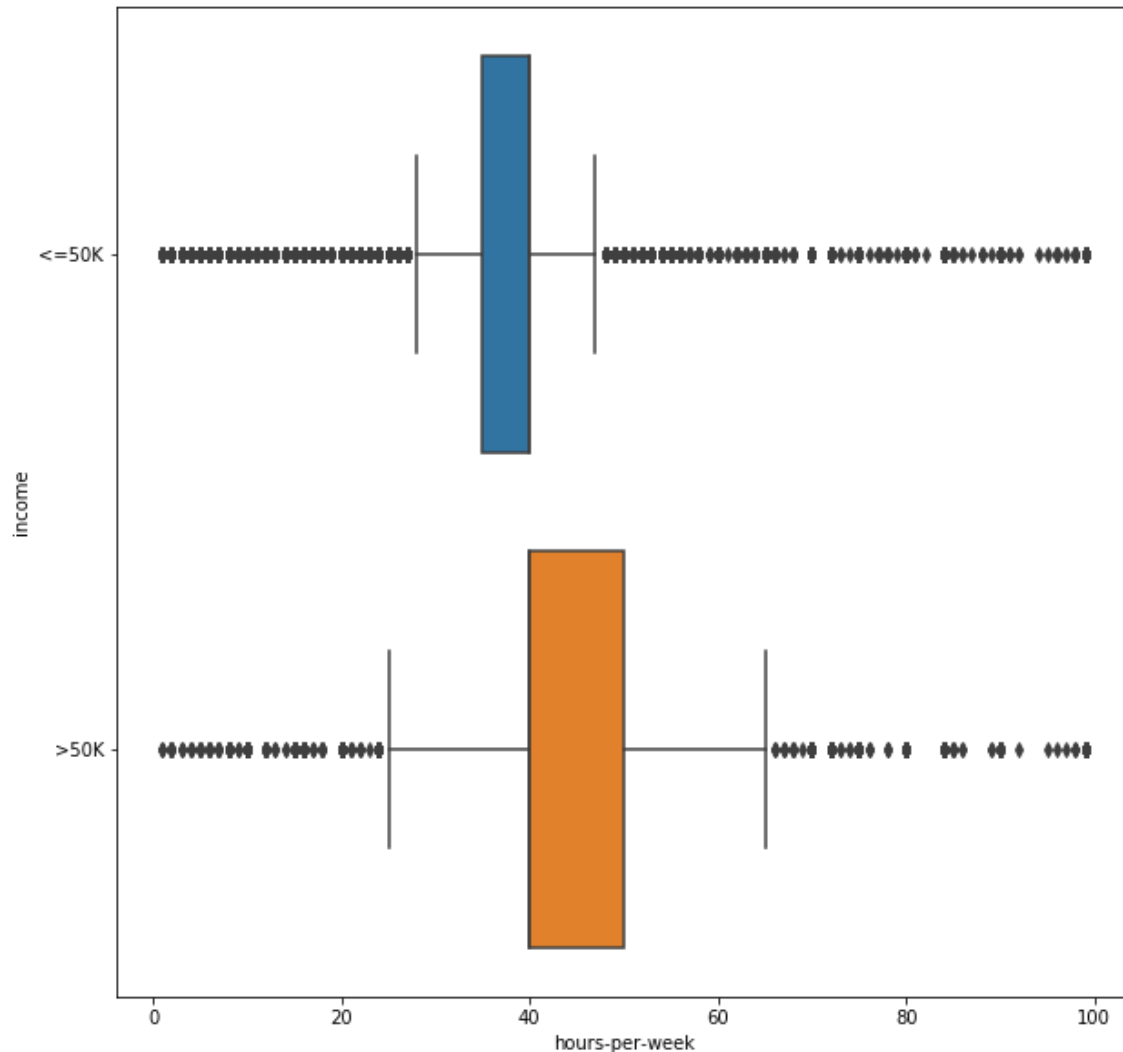
ii. Deskripsi

```
In [4]: # Jelaskan hasil diagram tersebut  
# histogram diatas dapat menjelaskan bahwa: atribut "income" termasuk simetris tidak stabi  
# dari 50 ribu) Upah minimum dan maksimum adalah lebih kecil dari sama dengan 50 dan Lebih
```

2. Bivariate analysis

2.1 hours-per-week - income (Hubungan)

```
In [24]: # Buat diagram, hubungan hours-per-week dengan income
fig = plt.figure(figsize=(10,10))
sns.boxplot(x="hours-per-week", y="income", data=data)
plt.show()
```



ii. Deskripsi

```
In [6]: # jelaskan hasil diagram tersebut
# histogram diatas dapat menjelaskan bahwa: atribut "hours per week - income" termasuk sim
# kurang dari sama dengan 50 ribu dalam kurun waktu 38 jam per minggu. Upah maksimum adalah
# kurun waktu 44 jam per minggu.
```

iii. Uji Hipotesis (untuk menguji hubungan antara hours-per-week & Income)

Dalam contoh ini kita dapat mengatakan bahwa:

Null Hypothesis :- tidak ada perbedaan Rerata kelompok pendapatan >50k dan kelompok pendapatan <=50k.

Alternate Hypothesis :- ada perbedaan Rerata kelompok pendapatan >50k dan kelompok

```
In [25]: # Buat uji hipotesis untuk menentukan null atau alternate hypothesis
data[['hours-per-week', 'income']].groupby(['income'], as_index=False).mean().sort_values(b
```

Out[25]:

	income	hours-per-week
1	>50K	45.468863
0	<=50K	38.892967

iv. Kesimpulan akhir

```
In [8]: # Berikan kesimpulan
# Disimpulkan bahwa terdapat perbedaan yang signifikan pada rerata jam kerja perminggu pada
# berpenghasilan <=50K. Hal Ini membuktikan bahwa jam kerja mingguan memiliki peranan dalam
# paling besar.
```