

Tugas 2 - Data Exploration

Anggota kelompok:

1. Muhammad Bintang Bahy (17/412643/PA/17962)
2. Ganjar Muhammad Parikesit (17/409434/PA/17741)
3. Perdo Kurniawan (17/412649/PA/17968)

Link video:

<https://github.com/bahybintang/rpld/raw/master/Income/tugas2.mkv>

1. Dataset

Dataset yang digunakan:

<https://www.kaggle.com/lodetomasi1995/income-classification>

Karena ketika kami coba menggunakan URL langsung dari Kaggle tidak bisa, akhirnya dataset kami upload ke GitHub, dan URL kami ambil dari GitHub.

URL:

https://raw.githubusercontent.com/bahybintang/rpld/master/Income/income_evaluation.csv

2. Feature

Feature yang digunakan adalah

1. age (numeric)
Merupakan umur dari orang yang terdapat di dataset.
2. education (categorical)
Tingkat pendidikan dari orang yang terdapat di dataset.
3. sex (categorical)
Jenis kelamin dari orang yang terdapat di dataset.
4. hours-per-week (categorical)
Jam kerja per minggu dari orang yang terdapat di dataset.
5. income (categorical/target)

Pendapatan per tahun dalam bentuk lebih besar dari US\$ 50.000,00 atau lebih kecil dari US\$ 50.000,00 dari orang yang terdapat di dataset.

File

File: income_evaluation.csv

URL: https://raw.githubusercontent.com/bahyubintang/rpld/master/Income/income_evaluation.csv

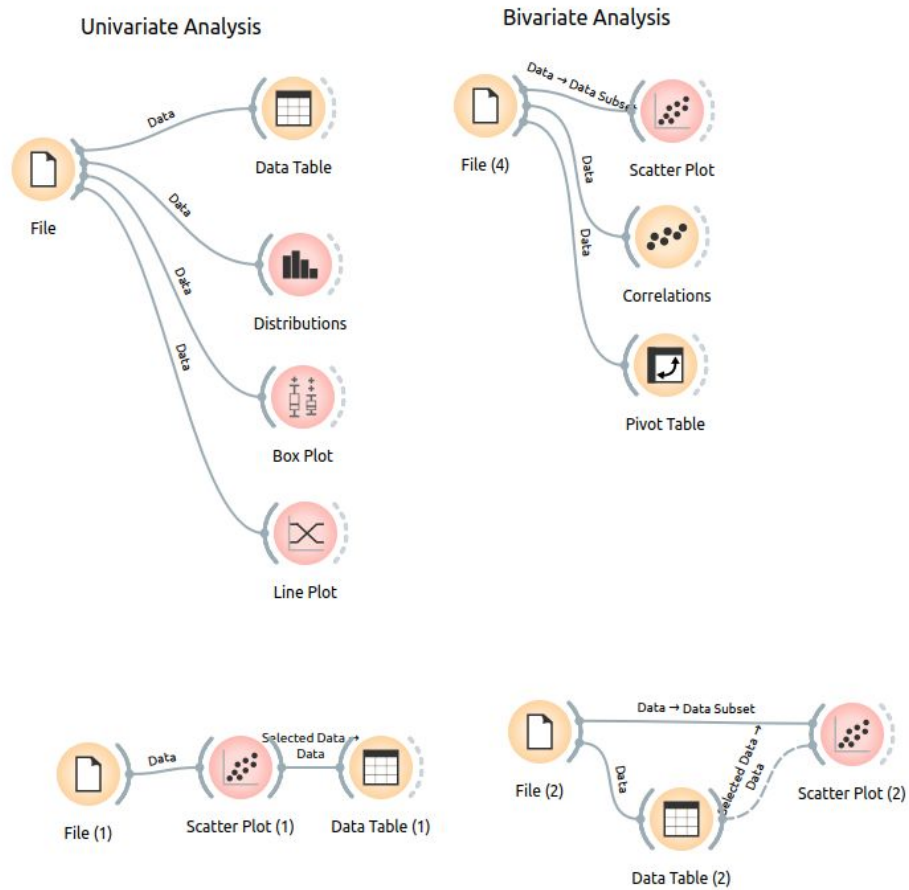
Info

32561 instance(s)
15 feature(s) (0.0% missing values)
Data has no target variable.
0 meta attribute(s)

Columns (Double click to edit)

	Name	Type	Role	Values
1	age	numeric	feature	
2	workclass	categorical	skip	Federal-gov, Local-gov, Never-worked, Private, Self-emp-inc, Self-emp-not-inc, State-gov, Without-pay
3	fnlwgt	text	skip	
4	education	categorical	feature	1st-4th, 5th-6th, 7th-8th, 9th, 10th, 11th, 12th, Assoc-acdm, Assoc-voc, Bachelors, Doctorate, HS-grad, Masters, Preschool, Prof-school, Some-college
5	education-num	numeric	skip	
6	marital-status	categorical	skip	Divorced, Married-AF-spouse, Married-civ-spouse, Married-spouse-absent, Never-married, Separated, Widowed
7	occupation	categorical	skip	Adm-clerical, Armed-forces, Craft-repair, Exec-managerial, Farming-fishing, Handlers-cleaners, Machine-op-inspct, Other-service, Priv-house-serv, Prof-specialty, Protective-serv, Sales, Tech-support, ...
8	relationship	categorical	skip	Husband, Not-in-family, Other-relative, Own-child, Unmarried, Wife
9	race	text	skip	Amer-indian-eskimo, Asian-pac-islander, Black, Other, White
10	sex	categorical	feature	Female, Male
11	capital-gain	numeric	skip	
12	capital-loss	numeric	skip	
13	hours-per-week	numeric	feature	
14	native-country	categorical	skip	Cambodia, Canada, China, Columbia, Cuba, Dominican-Republic, Ecuador, El-Salvador, England, France, Germany, Greece, Guatemala, Haiti, Holand-Netherlands, Honduras, Hong, Hungary, India, Iran, ...
15	income	categorical	target	<=50K, >50K

3. Workflow



Workflow yang kami buat, kami bagi menjadi empat sesuai dengan contoh yang ada di slides yang ada di course KS-RPLD di eLOK.

4. Univariate Analysis

4.1. Data Table

Data Table

Info

- 32561 instances (no missing data)
- 4 features
- Target with 2 values
- No meta attributes

Variables

- ☒ Show variable labels (if present)
- ☐ Visualize numeric values
- ☒ Color by instance classes

Selection

- ☒ Select full rows

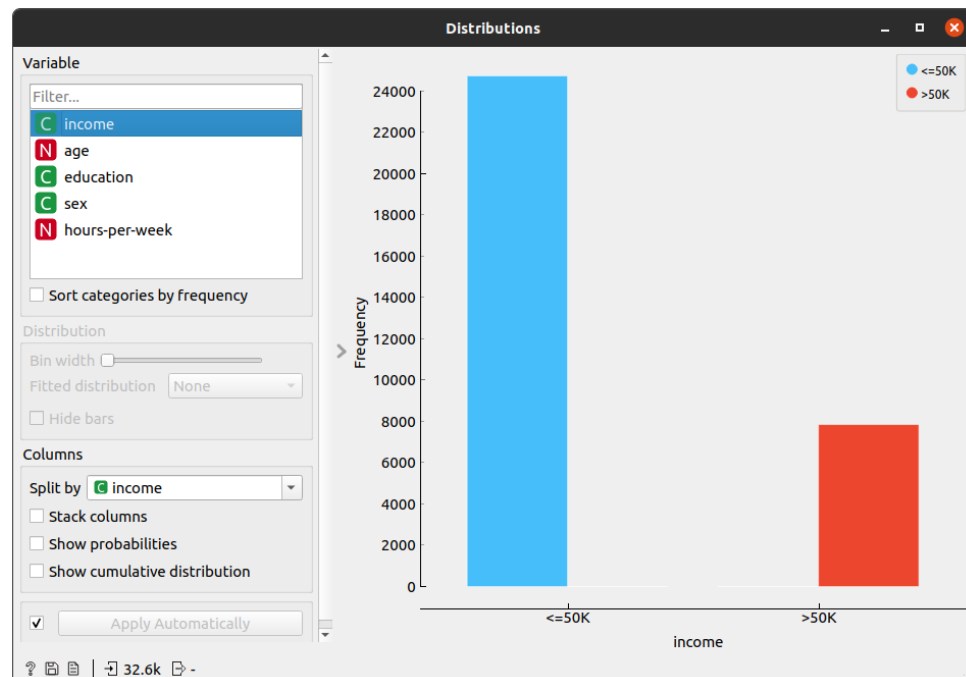
Restore Original Order

☒ Send Automatically

	income	age	education	sex	hours-per-week
1	<=50K	39	Bachelors	Male	40
2	<=50K	50	Bachelors	Male	13
3	<=50K	38	HS-grad	Male	40
4	<=50K	53	11th	Male	40
5	<=50K	28	Bachelors	Female	40
6	<=50K	37	Masters	Female	40
7	<=50K	49	9th	Female	16
8	>50K	52	HS-grad	Male	45
9	>50K	31	Masters	Female	50
10	>50K	42	Bachelors	Male	40
11	>50K	37	Some-college	Male	80
12	>50K	30	Bachelors	Male	40
13	<=50K	23	Bachelors	Female	30
14	<=50K	32	Assoc-acdm	Male	50
15	>50K	40	Assoc-voc	Male	40
16	<=50K	34	7th-8th	Male	45
17	<=50K	25	HS-grad	Male	35
18	<=50K	32	HS-grad	Male	40
19	<=50K	38	11th	Male	50
20	>50K	43	Masters	Female	45
21	>50K	40	Doctorate	Male	60
22	<=50K	54	HS-grad	Female	20
23	<=50K	35	9th	Male	40
24	<=50K	43	11th	Male	40
25	<=50K	59	HS-grad	Female	40
26	>50K	56	Bachelors	Male	40
27	<=50K	19	HS-grad	Male	40

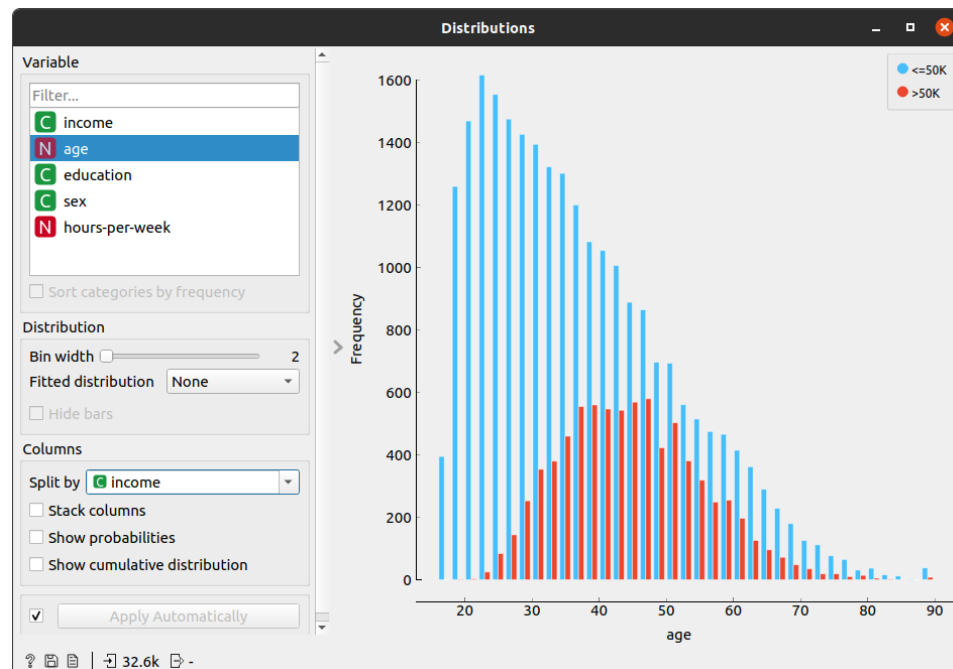
Jumlah data yang ada pada dataset adalah 32561 instance, dengan 4 fitur dan satu target yang memiliki nilai biner.

4.2. Distribution

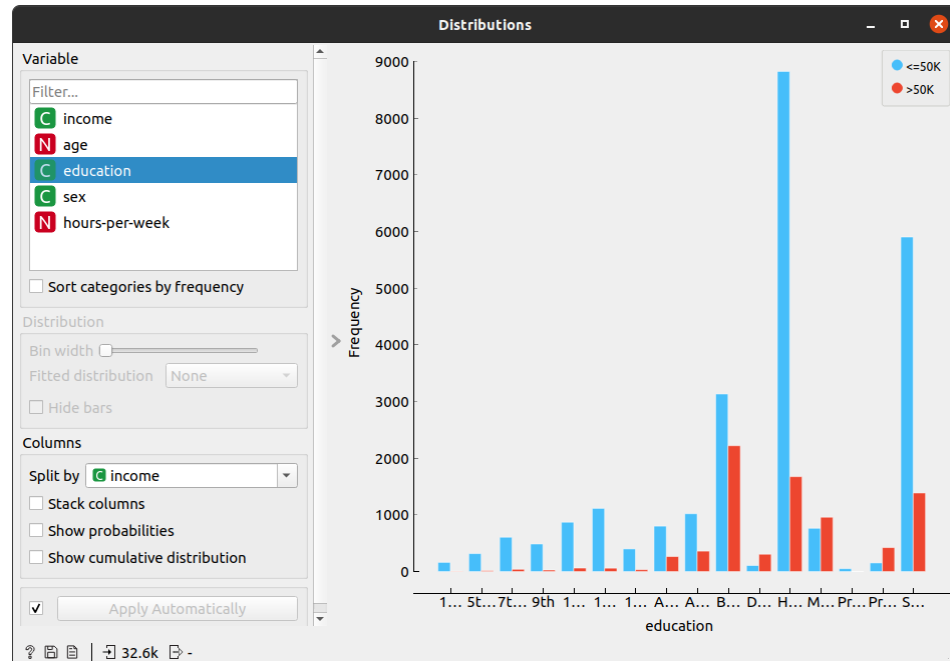


Pada visualisasi di atas dipilih variable dengan filter income lalu split by income untuk memisahkan label income >50K dan <=50K. Dari visualisasi di atas

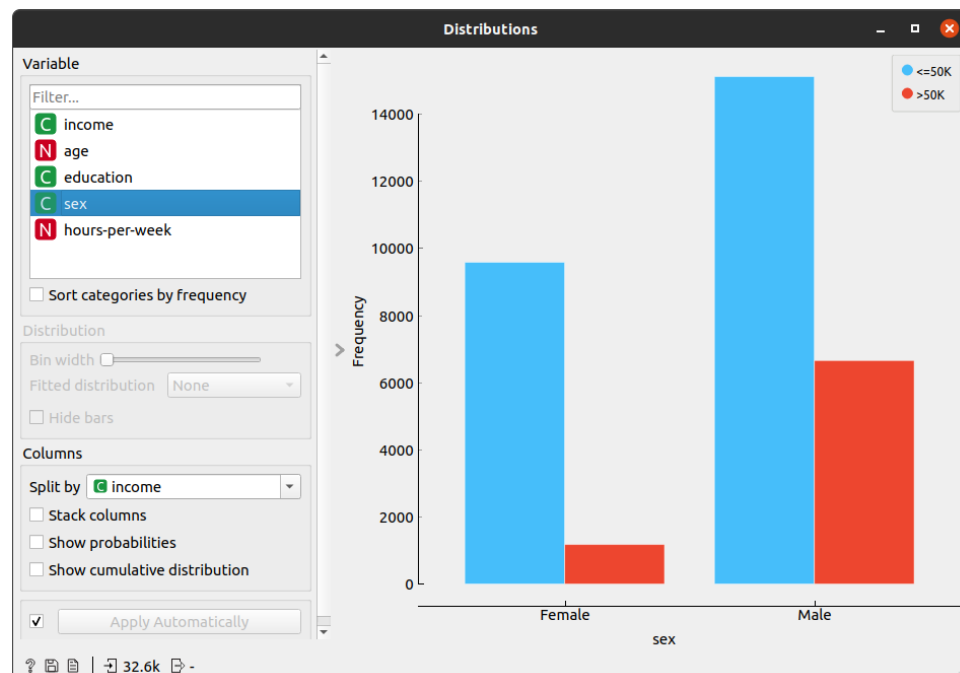
didapatkan informasi berupa data dengan income lebih dari 50K jauh lebih sedikit dibanding dengan income yang kurang dari atau sama dengan 50K.



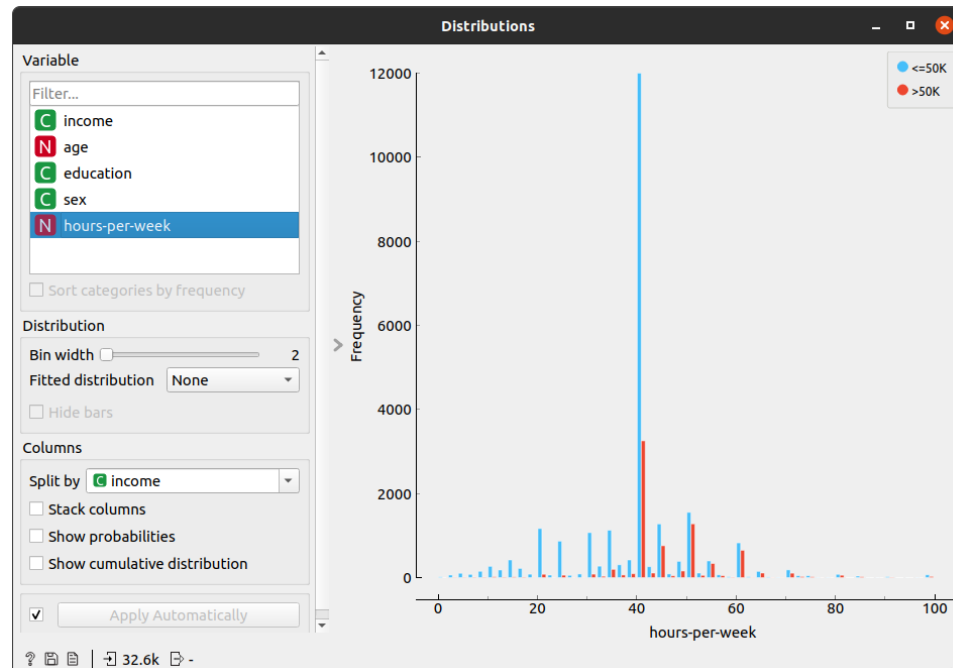
Pada visualisasi di atas kami memilih filter age lalu mengatur split by income. Dari visualisasi di atas didapatkan informasi berupa income $\leq 50K$ berjumlah lebih dari 400 data yang berkisar di antara umur 20 tahun sampai sekitar 60 tahun. Serta income yang lebih besar dari 50K tidak pernah lebih dari 600 data pada setiap usia yang berkisar antara 20 sampai 90 tahun.



Pada visualisasi di atas kami memilih Variabel filter education lalu split by income. Dari informasi di atas didapatkan bahwa income $\leq 50K$ dengan education HS-Grade memiliki data yang paling dari education lain dengan income $\leq 50K$. Serta income $> 50K$ dengan education Bachelor memiliki data yang paling banyak dari education lain dengan income $> 50K$.

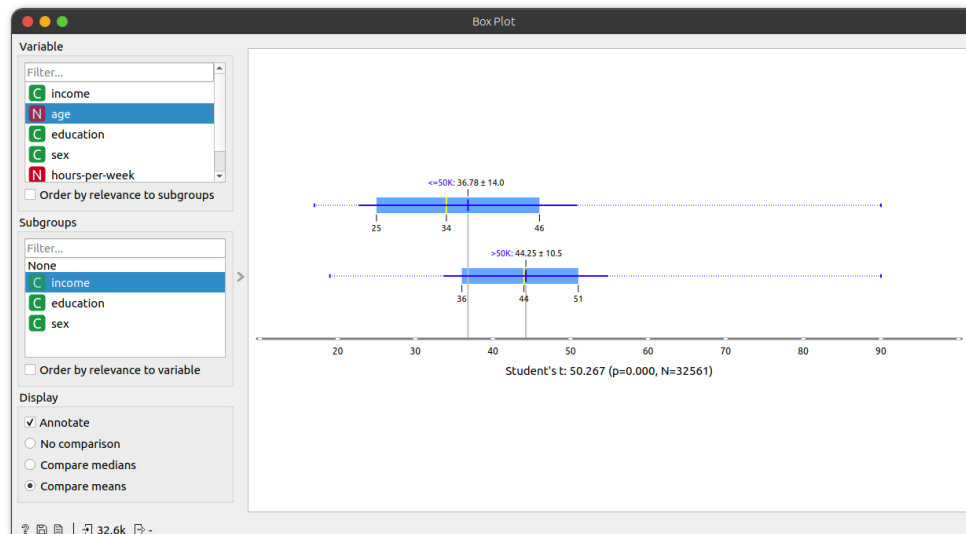
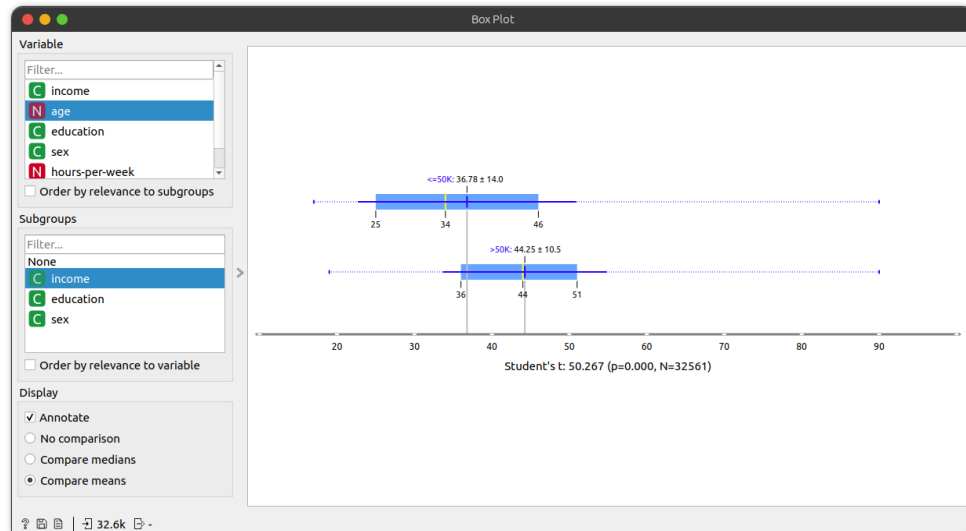


Dari visualisasi di atas kami memilih Variable filter sex serta split by income. Dari visualisasi tersebut didapatkan informasi berupa income $\leq 50K$ jauh lebih banyak dibanding $>50K$ pada kedua sex.



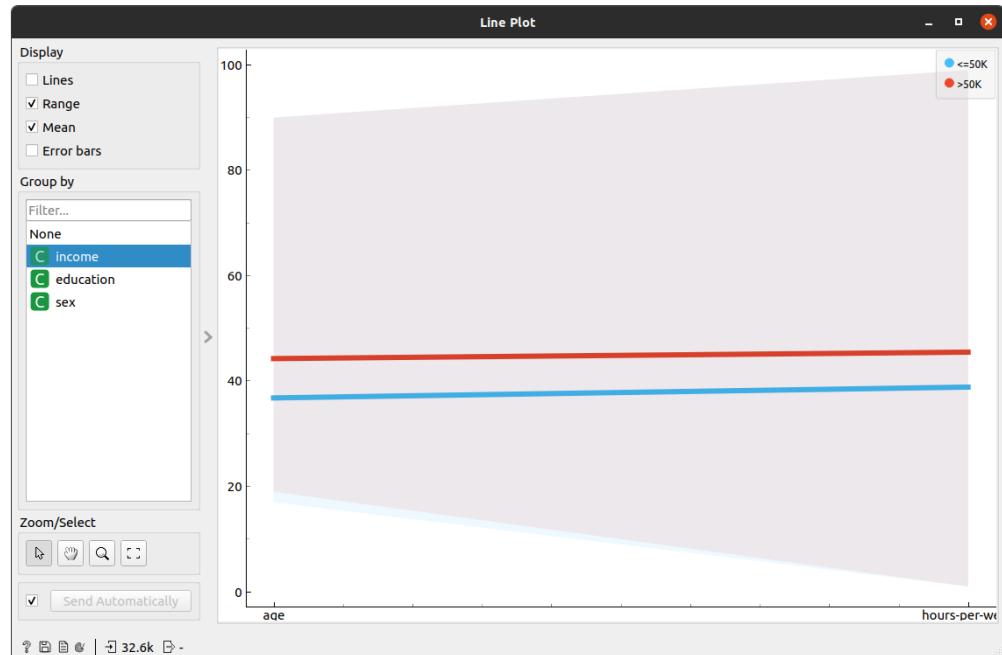
Pada widget Distribution dapat terlihat bagaimana distribusi data income, age, education, sex, dan hours-per-week pada dataset. Terlihat dataset cenderung lebih banyak data orang yang memiliki income lebih dari US\$ 50.000. Selain itu terlihat juga dataset cenderung lebih banyak data orang berjenis kelamin laki-laki.

4.3. Box Plot



Berikut adalah hasil BoxPlot dari dataset yang kami gunakan. Kami membandingkan dua fitur continuous dari dataset kami dengan target income. Dari hasil BoxPlot dapat dilihat bahwa untuk gaji lebih dari 50k maka ada kecenderungan bahwa jam kerja dan umur lebih besar dibandingkan dengan yang gajinya dibawah 50k.

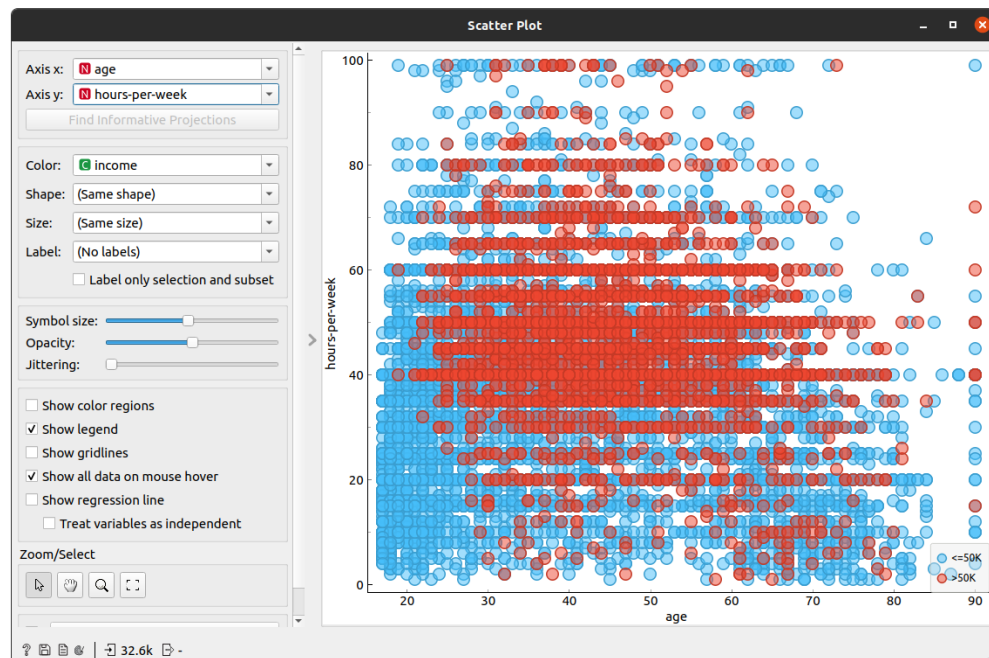
4.4. Line Plot



Diatas adalah line plot yang menunjukkan rata-rata umur dan jam kerja per-minggu untuk pendapatan diatas 50k dan dibawah 50k.

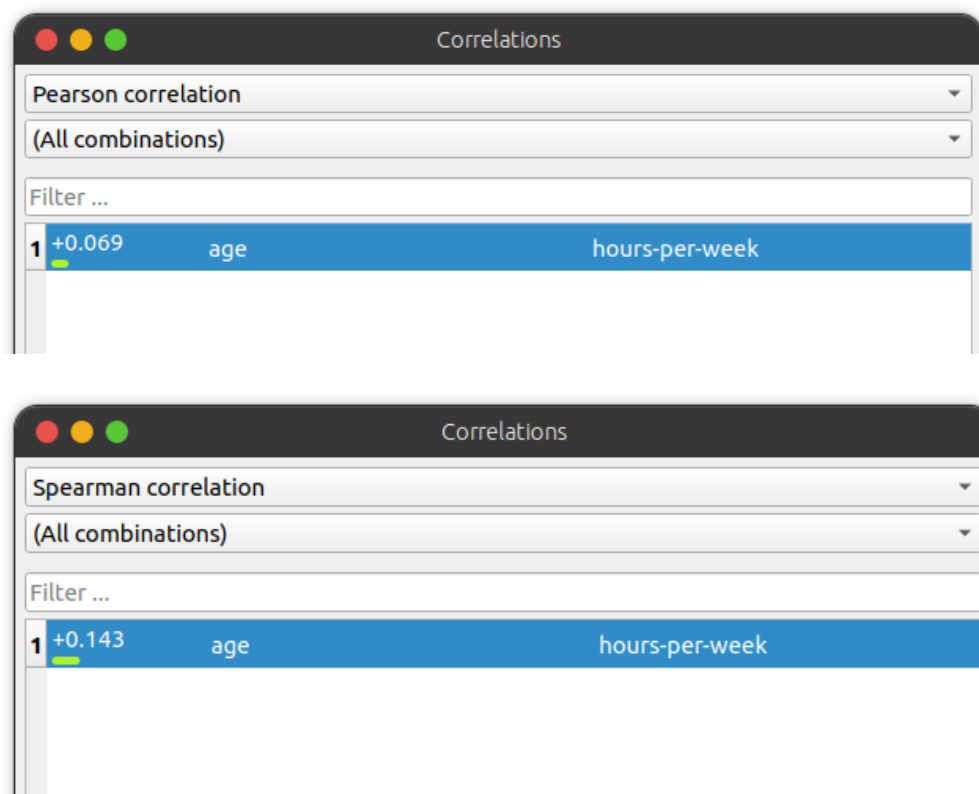
5. Bivariate Analysis

5.1. Scatter Plot



Diatas merupakan scatter plot dari dataset kami yang membandingkan dua fiturnya yaitu umur dan jam kerja tiap minggu, sedangkan warna merah menandakan gaji diatas 50k dan warna biru menandakan gaji dibawah 50k. Dari visualisasi tersebut kami melihat bahwa tidak ada data yang umurnya rendah dan jam kerja tiap minggunya rendah tetapi memiliki gaji diatas 50k. Dan untuk data dengan gaji diatas 50k cenderung berada di nilai tengah jam kerja perminggu dan umur.

5.2. Correlations



Dari hasil perhitungan correlation menggunakan metode Spearman dan Pearson dapat dilihat bahwa ada korelasi positif antara umur dengan jam kerja per minggu. Yang berarti semakin tinggi umur cenderung semakin tinggi pula jam kerja per minggu, begitu pula sebaliknya.

5.3. Pivot Table

Pivot Table

Rows

C

sex

Columns

C

income

Values

N

age

Aggregations

☒ Count

☐ Count defined

☐ Sum

☐ Mean

☐ Var

☐ Median

☐ Majority

☐ Mode

☐ Min

☐ Max

sex

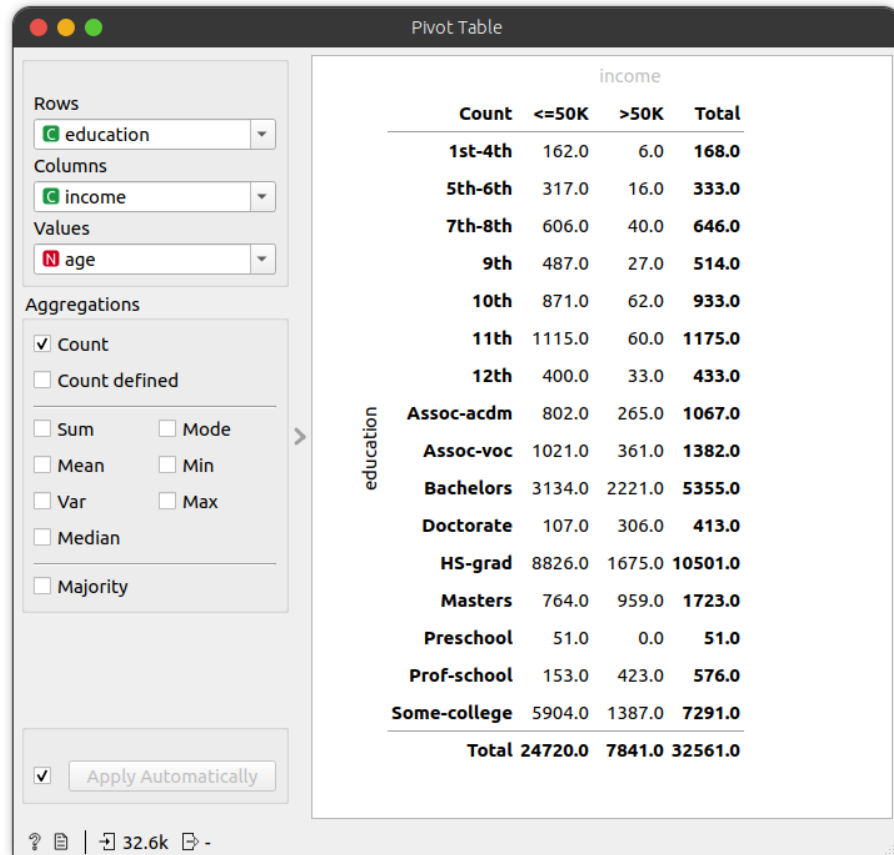
income

Count	<=50K	>50K	Total
Female	9592.0	1179.0	10771.0
Male	15128.0	6662.0	21790.0
Total	24720.0	7841.0	32561.0

?

32.6k

Dari pivot table perbandingan pendapatan dengan jenis kelamin, didapatkan bahwa untuk pendapatan dibawah 50k jumlah laki-laki lebih banyak. Tapi perlu diingat bahwa dataset dari awal memang lebih banyak laki-laki. Untuk pendapatan diatas 50k juga lebih banyak laki-laki, namun perbandingannya cukup besar sekitar 6:1, jauh diatas perbandingan laki-laki dengan perempuan pada keseluruhan dataset.



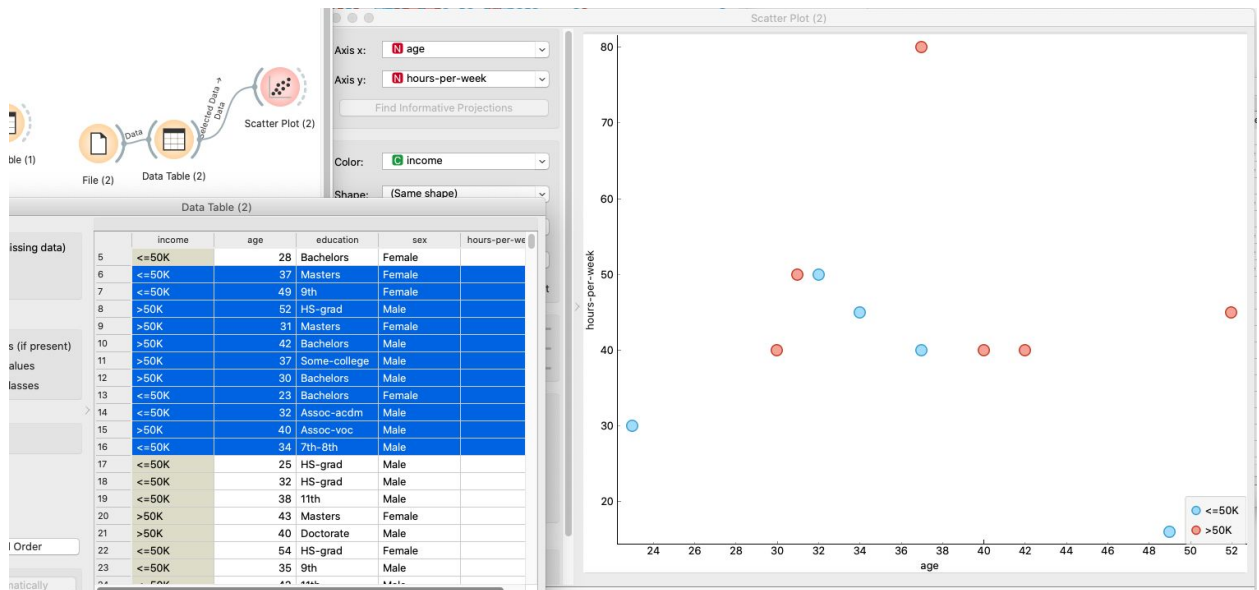
Dari pivot table perbandingan tingkat pendidikan dengan income, kami bisa melihat bahwa hampir di keseluruhan tingkat pendidikan pendapatan yang kurang dari 50k lebih banyak, kecuali untuk Masters dan Prof-school yang pendapatan lebih dari 50k lebih banyak.

6. Interactive Visualization



Berikut adalah percobaan kami dengan memilih 2 data pada widget Scatter Plot di sekitar umur 70 hingga 80 tahun dengan jam kerja mendekati 100, terlihat pada widget Data Table data yang kita pilih pada Scatter Plot

7. Visualization of Data Subset



Pada percobaan kami di Widget Data Table kami memilih beberapa data secara berurutan dan dapat kita lihat penyebarannya pada Widget Scatter Plot.

Daftar Pustaka

Kaggle.com. 2021. *Income classification*. [online] Available at:

<<https://www.kaggle.com/lodetomasi1995/income-classification>> [Accessed 26 February 2021].