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)

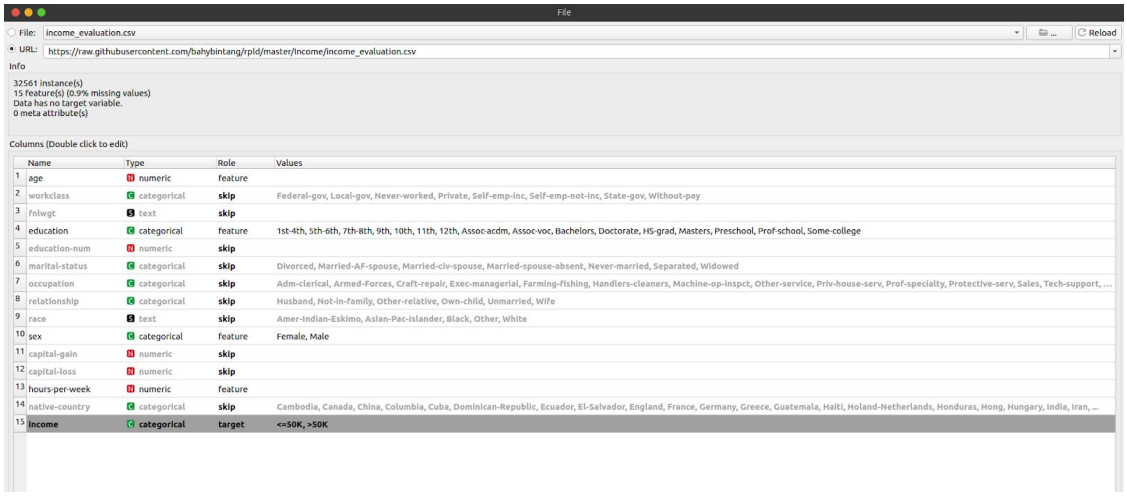
   Merupakan umur dari orang yang terdapat di dataset.
2. education (categorical)

   Tingkat pendidikan dari orang yang terdapat di dataset.
3. sex (categorical)

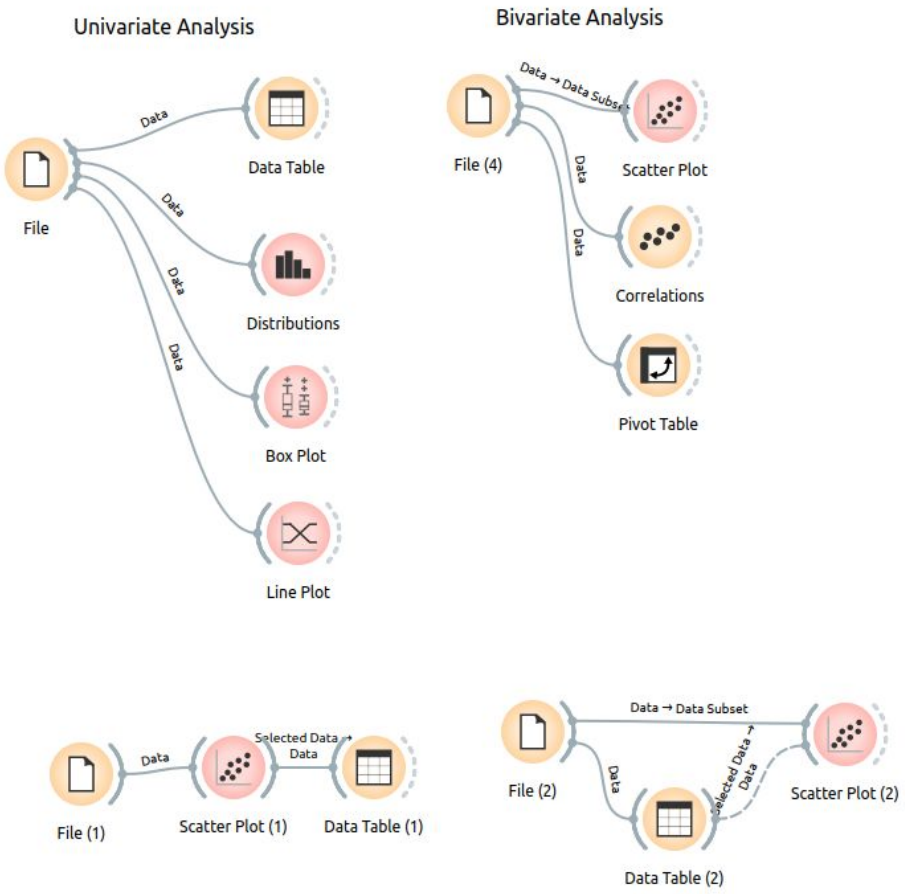
   Jenis kelamin dari orang yang terdapat di dataset.
4. hours-per-week (categorical)

   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.



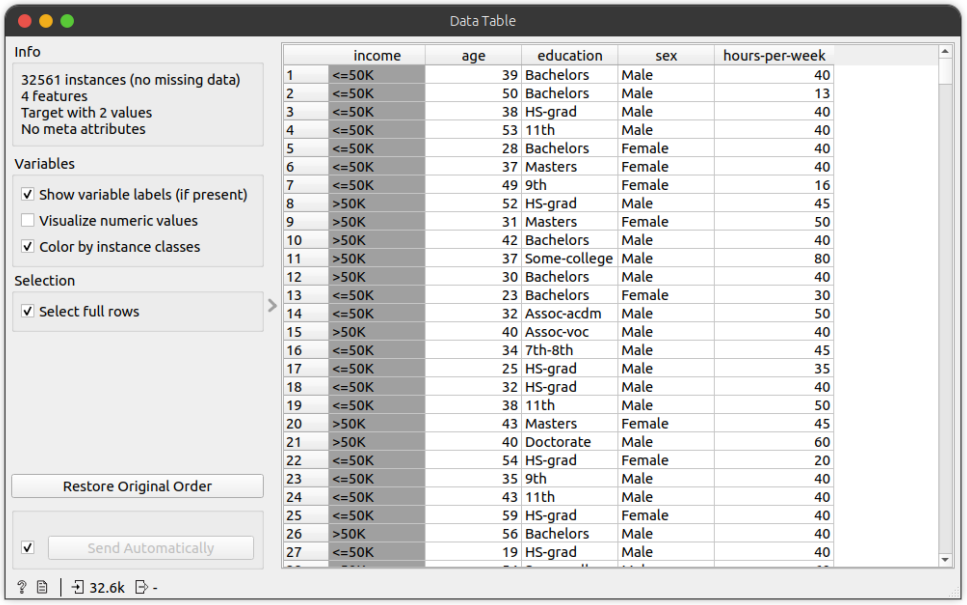
3. Workflow



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

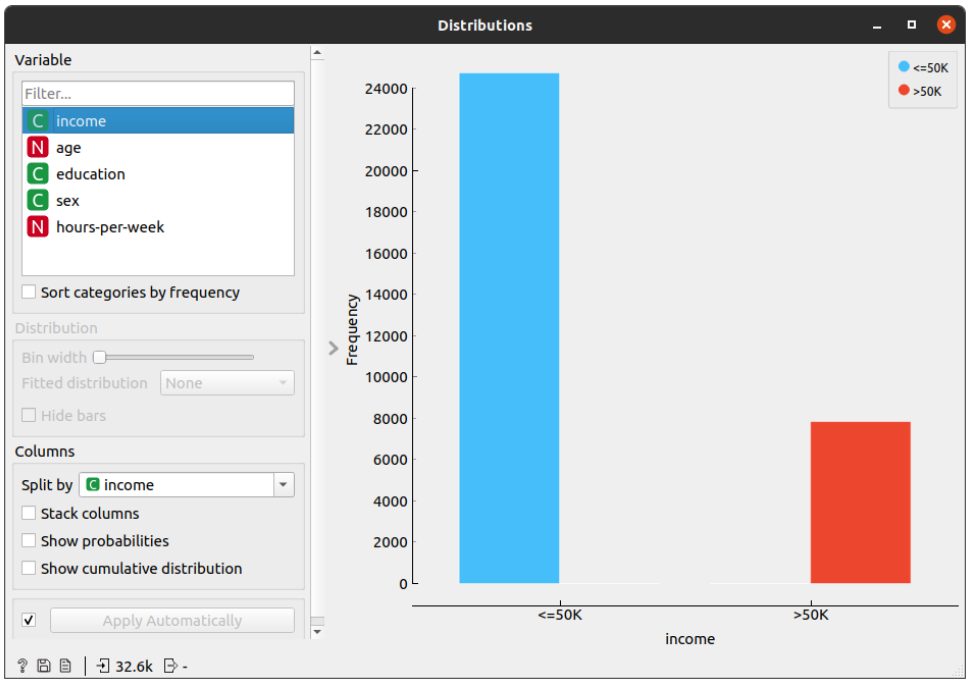
4. Univariate Analysis

4.1. Data Table



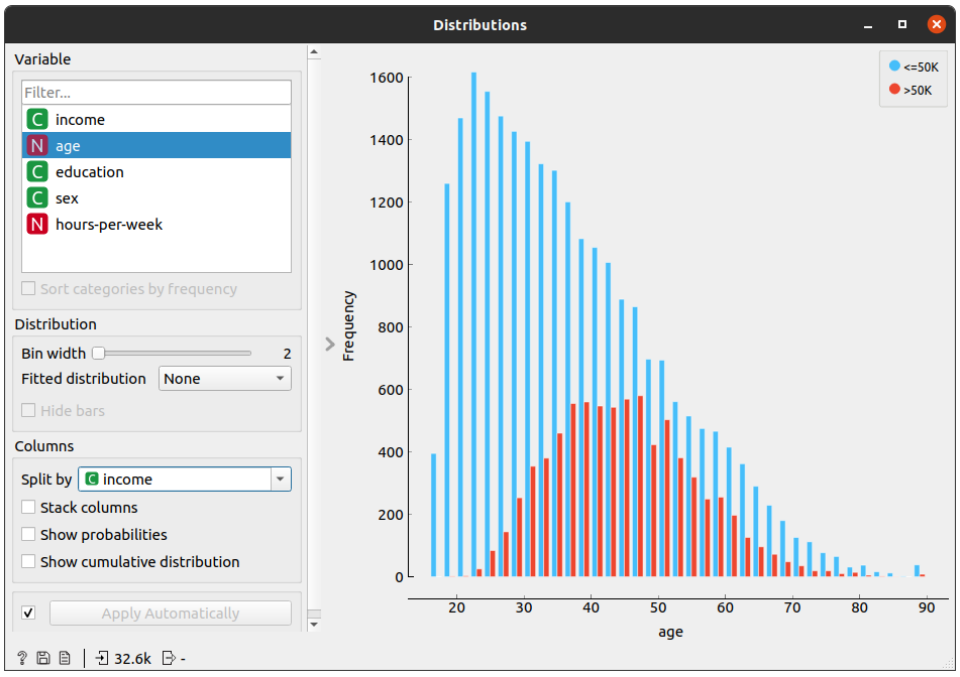
Jumlah data yang ada pada dataset adalah 32561 instance, dengan 4 fitur dan satu target yang memiliki nilai biner.

4.2. Distribution

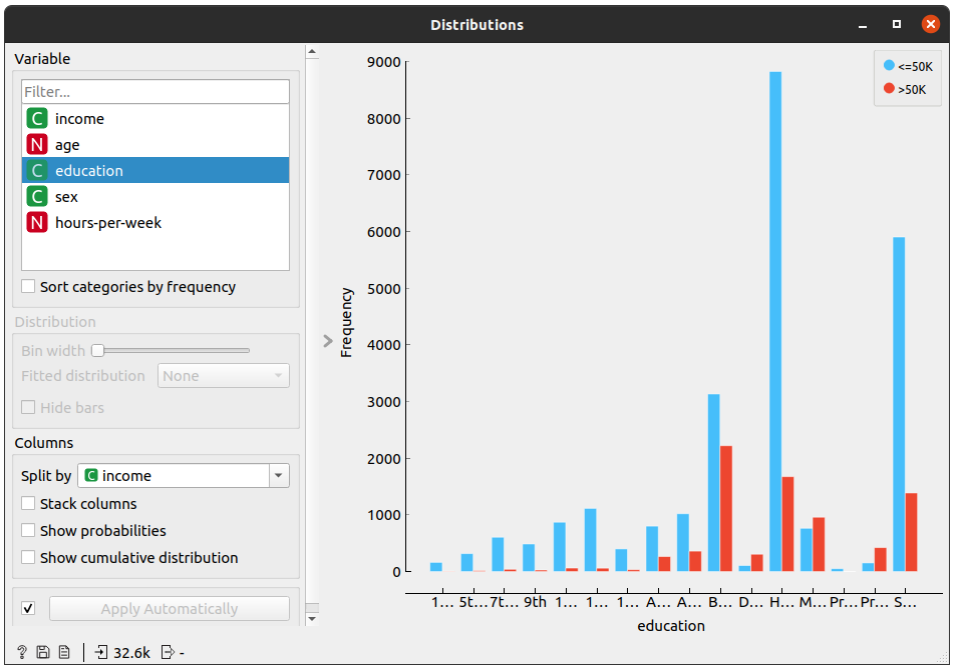


Pada visualisasi di atas dipilih variable dengan filter income lalu split by income untuk memisahkan label income >50K dan <=50K. Dari visualisasi di atas

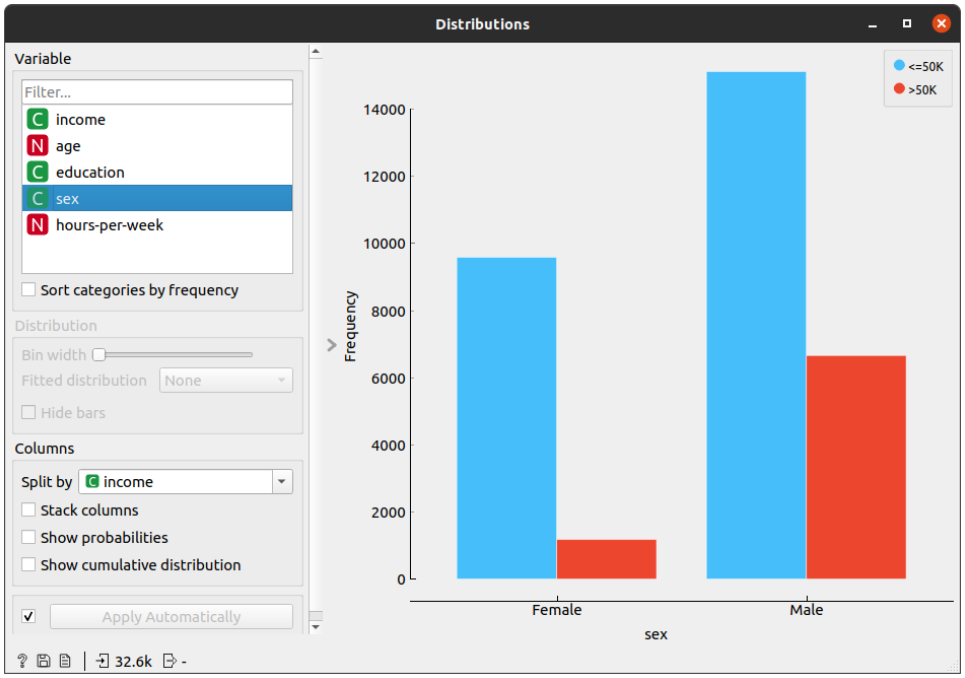
didapatkan informasi berupa data dengan income lebih dari 50K jauh lebih sedikit dibanding dengan income yang kurang dari atau sama dengan 50K.



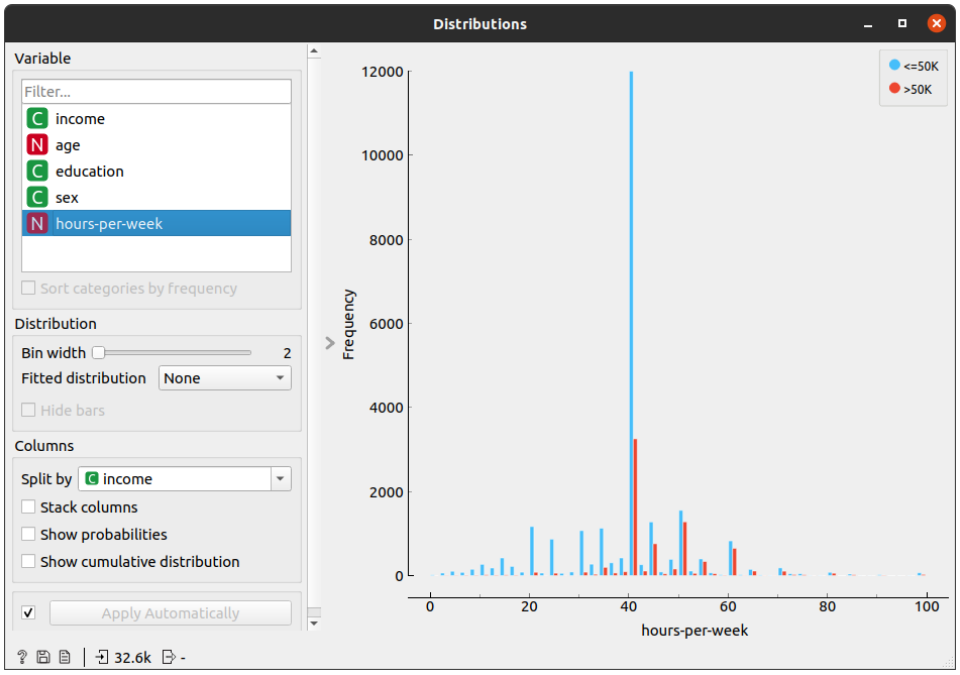
Pada visualisasi di atas kami memilih filter age lalu mengatur split by income. Dari visualisasi di atas didapatkan informasi berupa income <=50K berjumlah lebih dari 400 data yang berkisar di antara umur 20 tahun sampai sekitar 60 tahun. Serta income yang lebih besar dari 50K tidak pernah lebih dari 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 <=50K dengan education HS-Grade memiliki data yang paling dari education lain dengan income <=50K. Serta income >50K dengan education Bachelor memiliki data yang paling banyak dari education lain dengan income >50K.

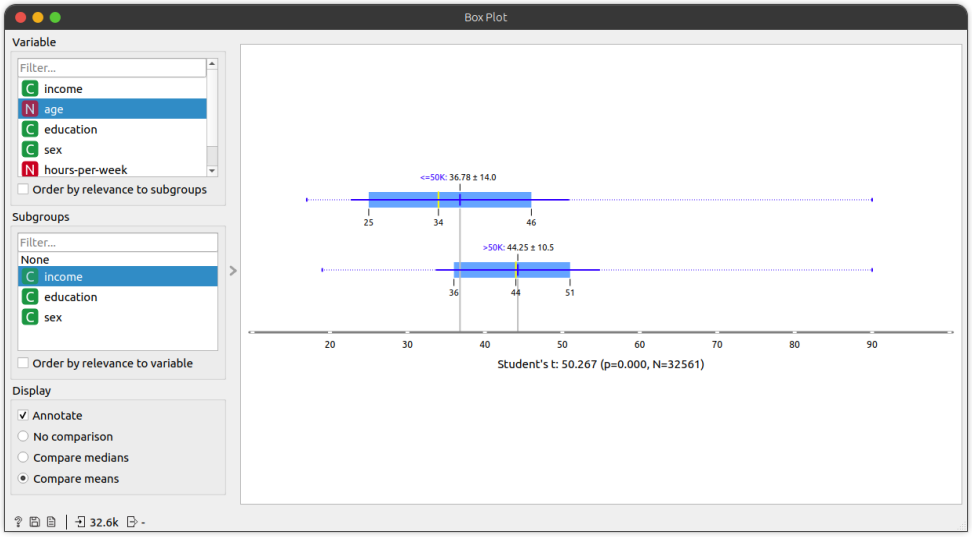


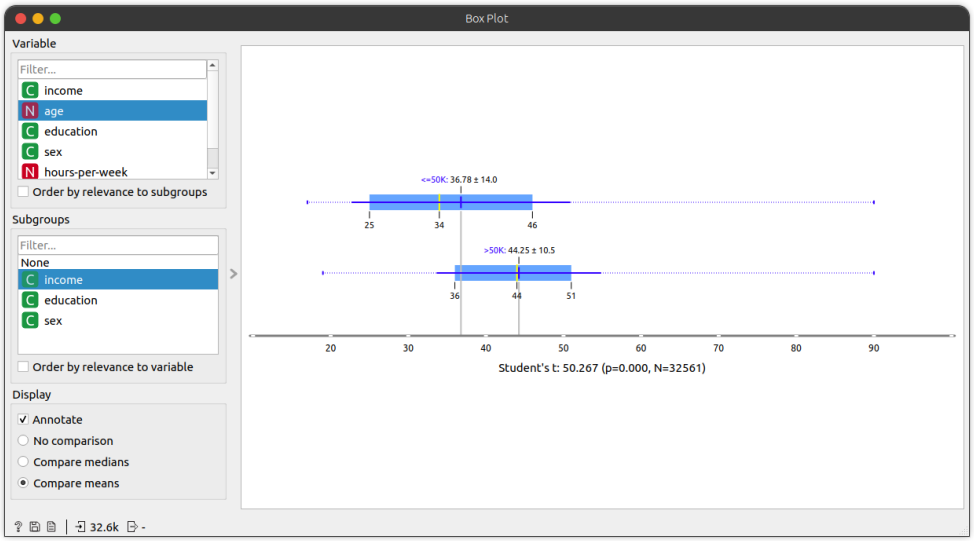
Dari visualisasi di atas kami memilih Variable filter sex serta split by income. Dari visualisasi tersebut didapatkan informasi berupa income <=50K jauh lebih banyak dibanding >50K pada kedua sex.



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.

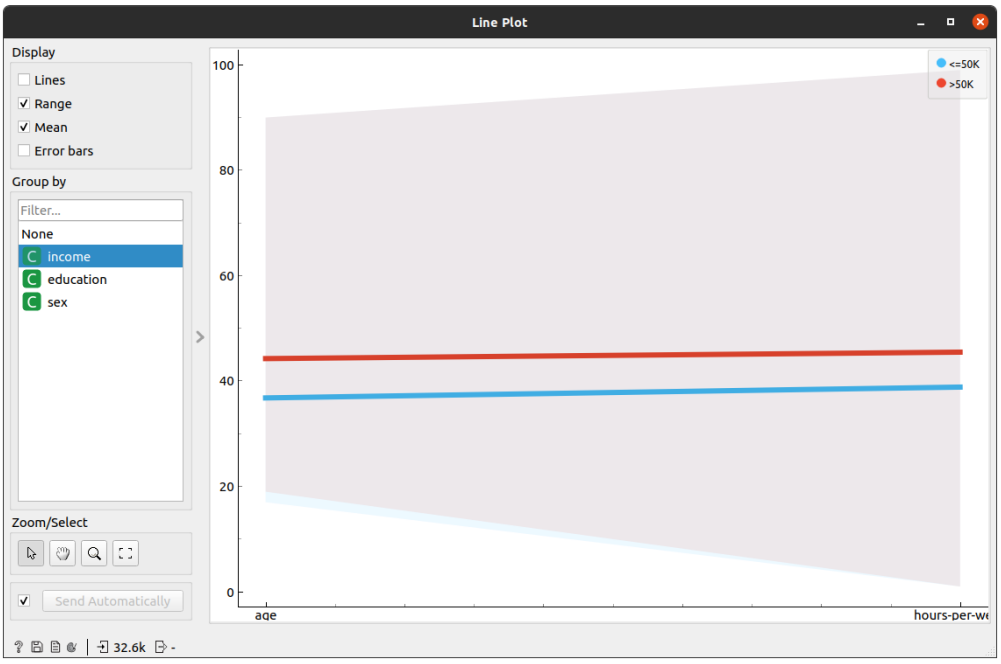
4.3. Box Plot





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



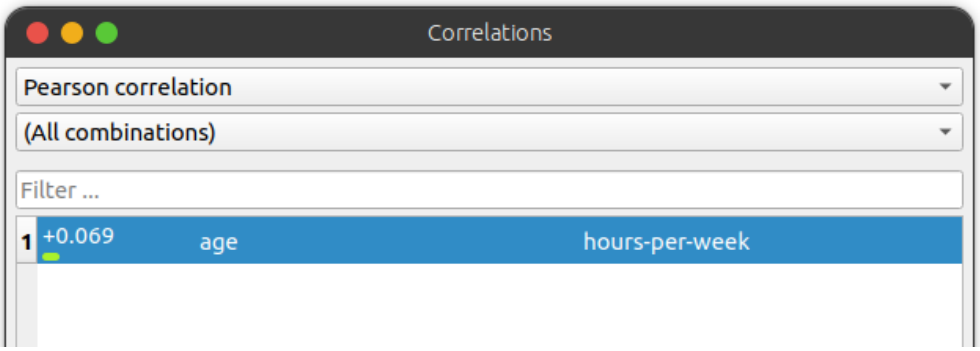
Diatas adalah line plot yang menunjukkan rata-rata umur dan jam kerja per-minggu untuk pendapatan diatas 50k dan dibawah 50k.

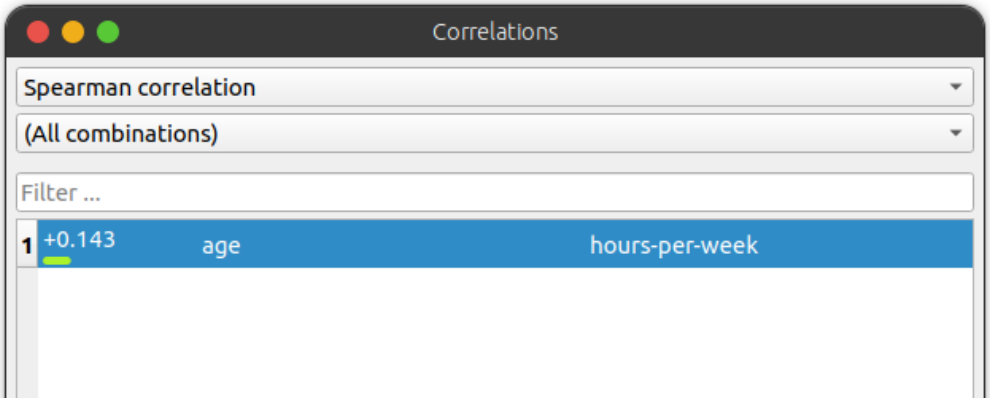
5. Bivariate Analysis

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.

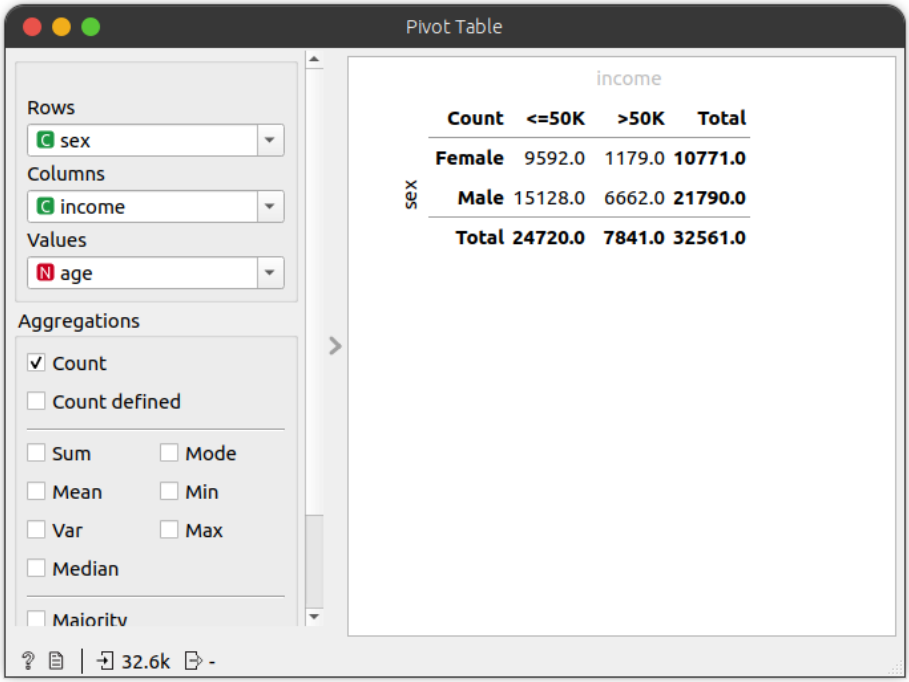
5.2. Correlations



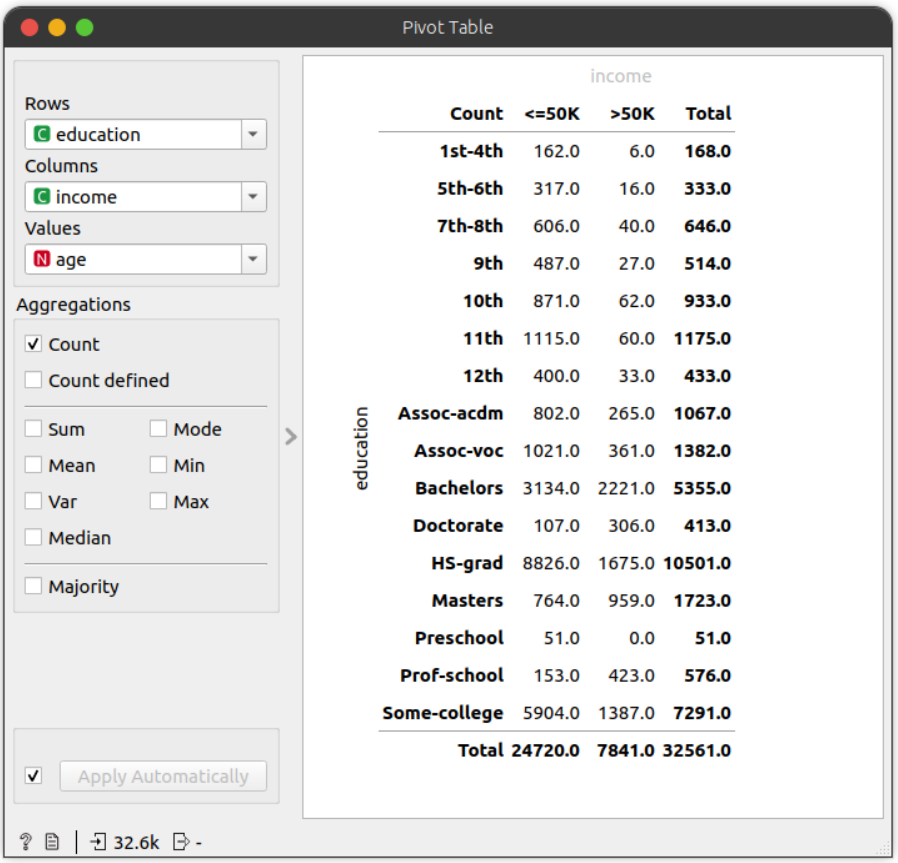


Dari hasil perhitungan correlation menggunakan metode Spearman dan Pearson dapat dilihat bahwa ada korelasi positif antara umur dengan jam kerja per minggu. Yang berarti semakin tinggi umur cenderung semakin tinggi pula jam kerja per minggu, begitu pula sebaliknya.

5.3. Pivot Table



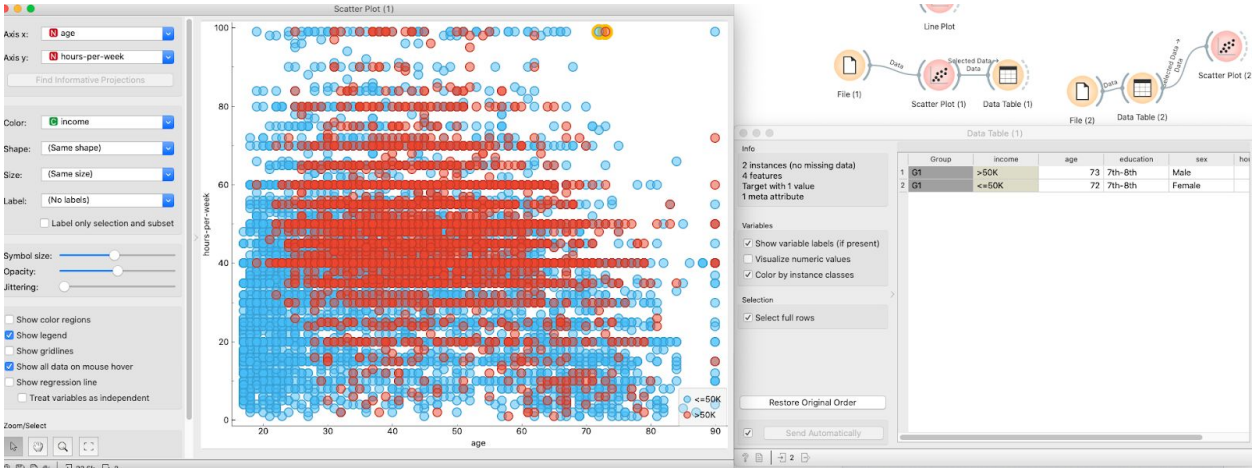
Dari pivot table perbandingan pendapatan dengan jenis kelamin, didapatkan bahwa untuk pendapatan dibawah 50k jumlah laki-laki lebih banyak. Tapi perlu diingat bahwa dataset dari awal memang lebih banyak laki-laki. Untuk pendapatan diatas 50k juga lebih banyak laki-laki, namun perbandingannya cukup besar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



education	Count	<=50K	>50K	Total
1st-4th		162.0	6.0	**168.0**
5th-6th		317.0	16.0	**333.0**
7th-8th		606.0	40.0	**646.0**
9th		487.0	27.0	**514.0**
10th		871.0	62.0	**933.0**
11th		1115.0	60.0	**1175.0**
12th		400.0	33.0	**433.0**
Assoc-acdm		802.0	265.0	**1067.0**
Assoc-voc		1021.0	361.0	**1382.0**
Bachelors		3134.0	2221.0	**5355.0**
Doctorate		107.0	306.0	**413.0**
HS-grad		8826.0	1675.0	**10501.0**
Masters		764.0	959.0	**1723.0**
Preschool		51.0	0.0	**51.0**
Prof-school		153.0	423.0	**576.0**
Some-college		5904.0	1387.0	**7291.0**
Total	**24720.0**	**7841.0**	**32561.0**	

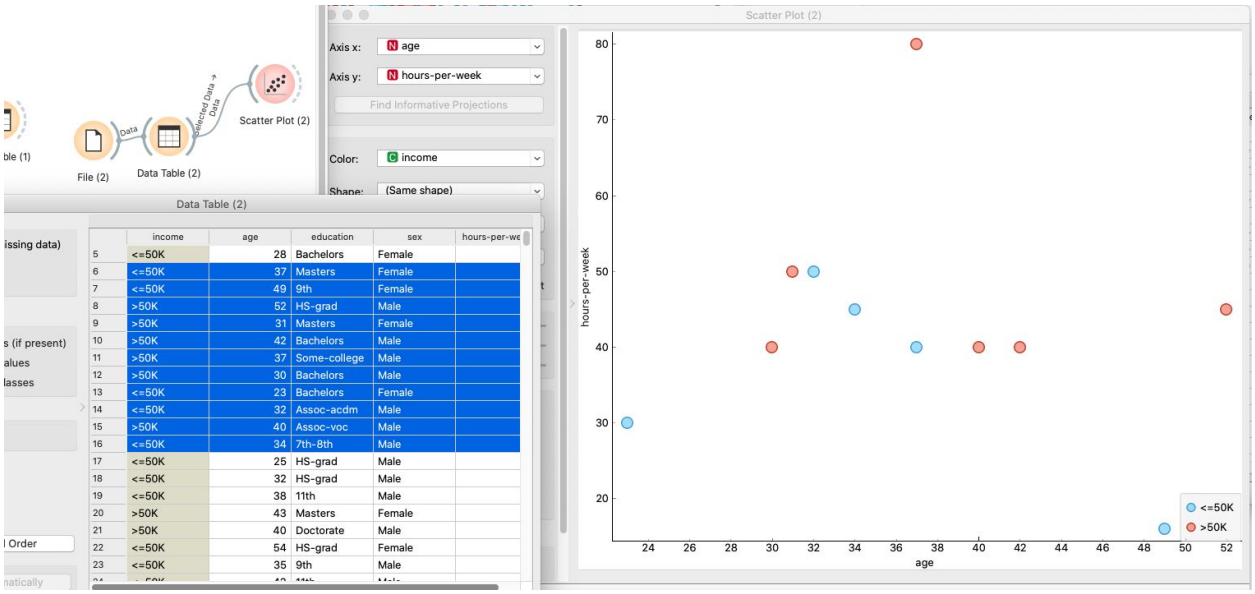
Dari pivot table perbandingan tingkat pendidikan dengan income, kami bisa melihat bahwa hampir di keseluruhan tingkat pendidikan pendapatan yang kurang dari 50k lebih banyak, kecuali untuk Masters dan Prof-school yang pendapatan lebih dari 50k lebih banyak.

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].