Базисни софтуерни продукти за работа с данни

Ангел Марчев, мл.

Калоян Харалампиев

Въведение

- Невъзможно е да се покрият всички съществуващи продукти

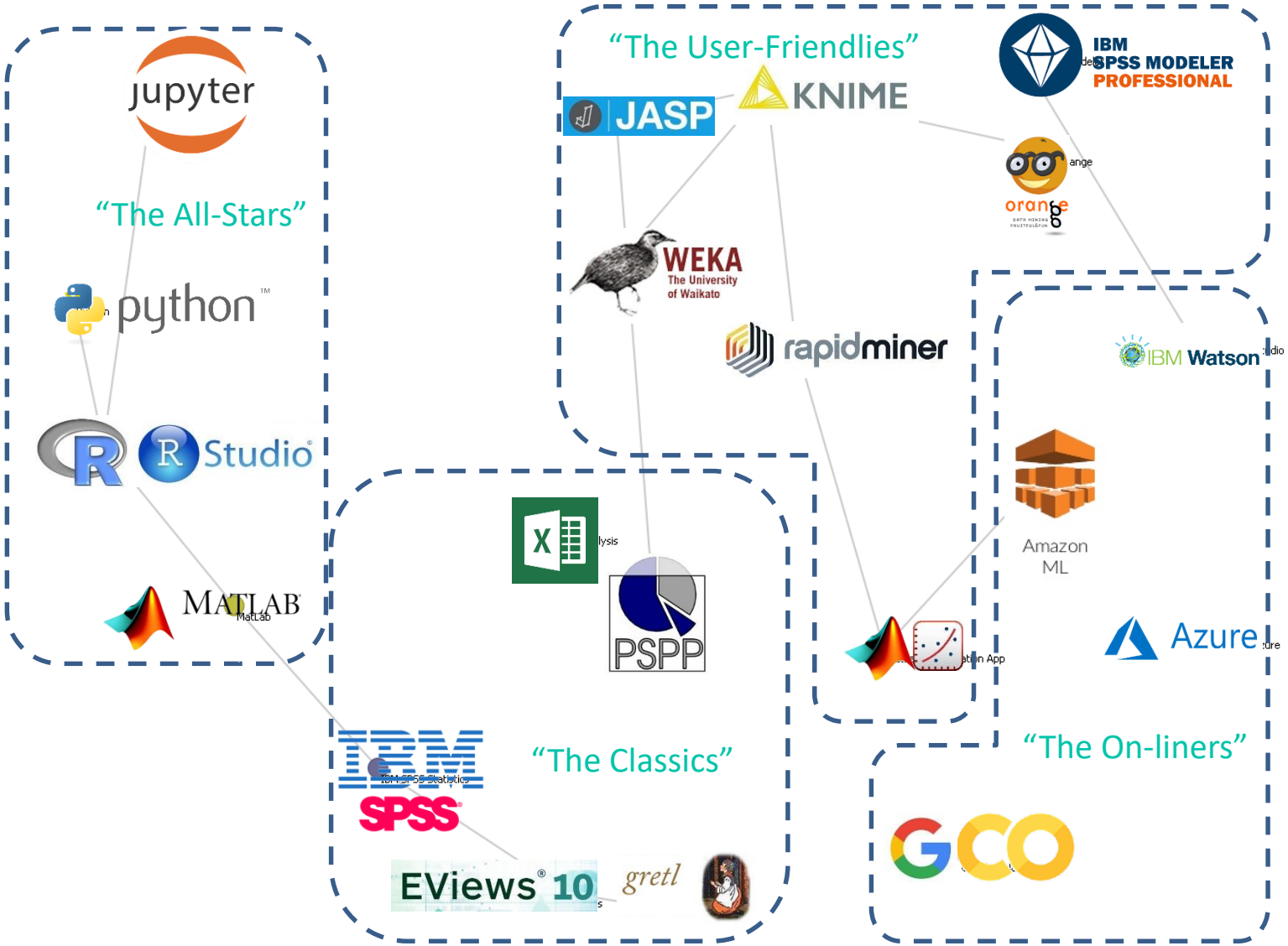
- Затова ние редуцираме броя на продуктите до тези, които самите ние използваме

- Но задачата все още е трудна, защото:
  - Има различни типове продукти
  - Всеки продукт може да се класифицира по много критерии
  - Всеки продукт може да се използва за много цели

Критерии за класификация

- Приложение
  - Статистика
  - Иконометрия
  - Data mining
- Начин на работа
  - Писане на програмен код
  - Работа с менюта и прозорци
  - Работа с графични обекти
  - Работа онлайн
- Цена
  - Безплатни
  - Платени
- Свързаност (приемственост)
- Популярност
- Интерактивност

Една възможна класификация

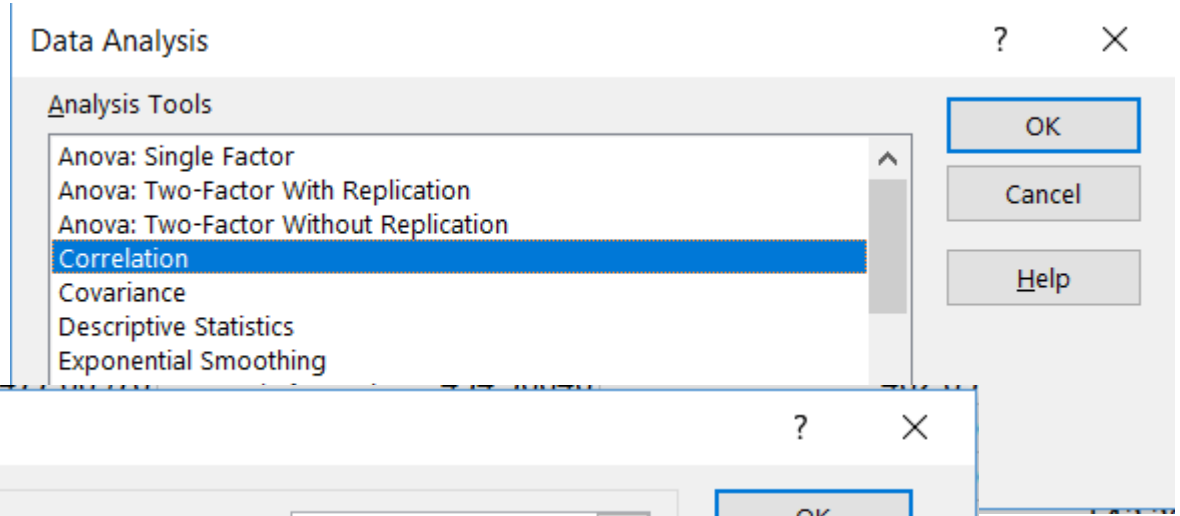


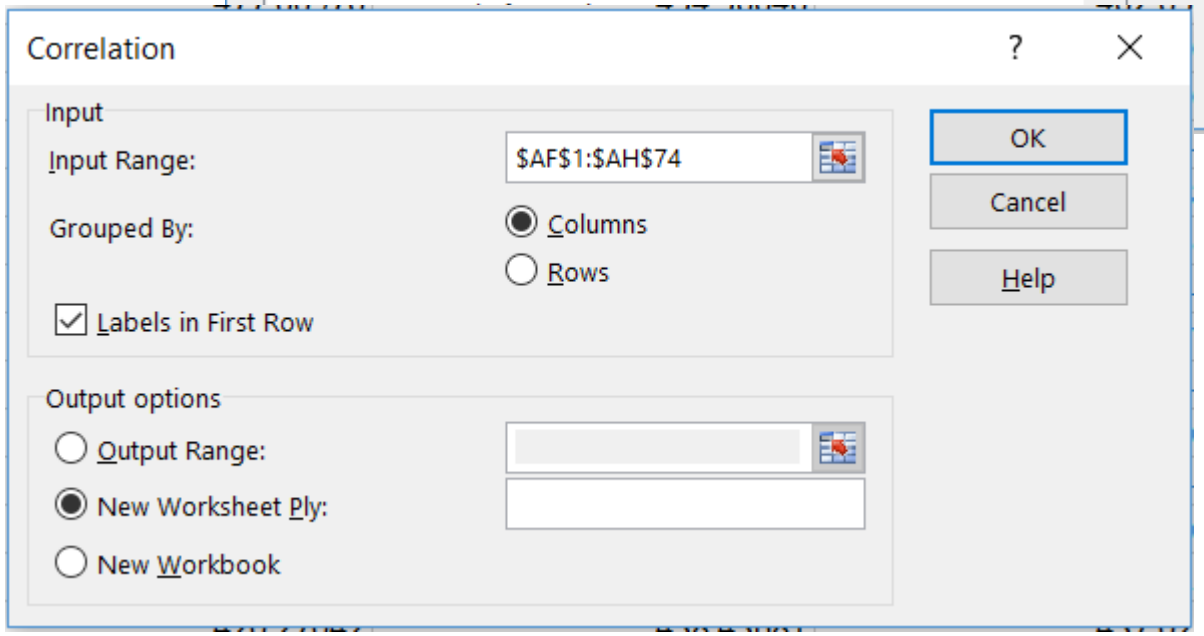
The Classics

Excel Data Analysis

- Приложение: Статистика
- Начин на работа: Менюта и прозорци
- Цена: Платен

- Предимства: Достъпност (почти всеки има Excel)
- Недостатъци: Работи се чрез избор на клетки, а не чрез имената на променливите



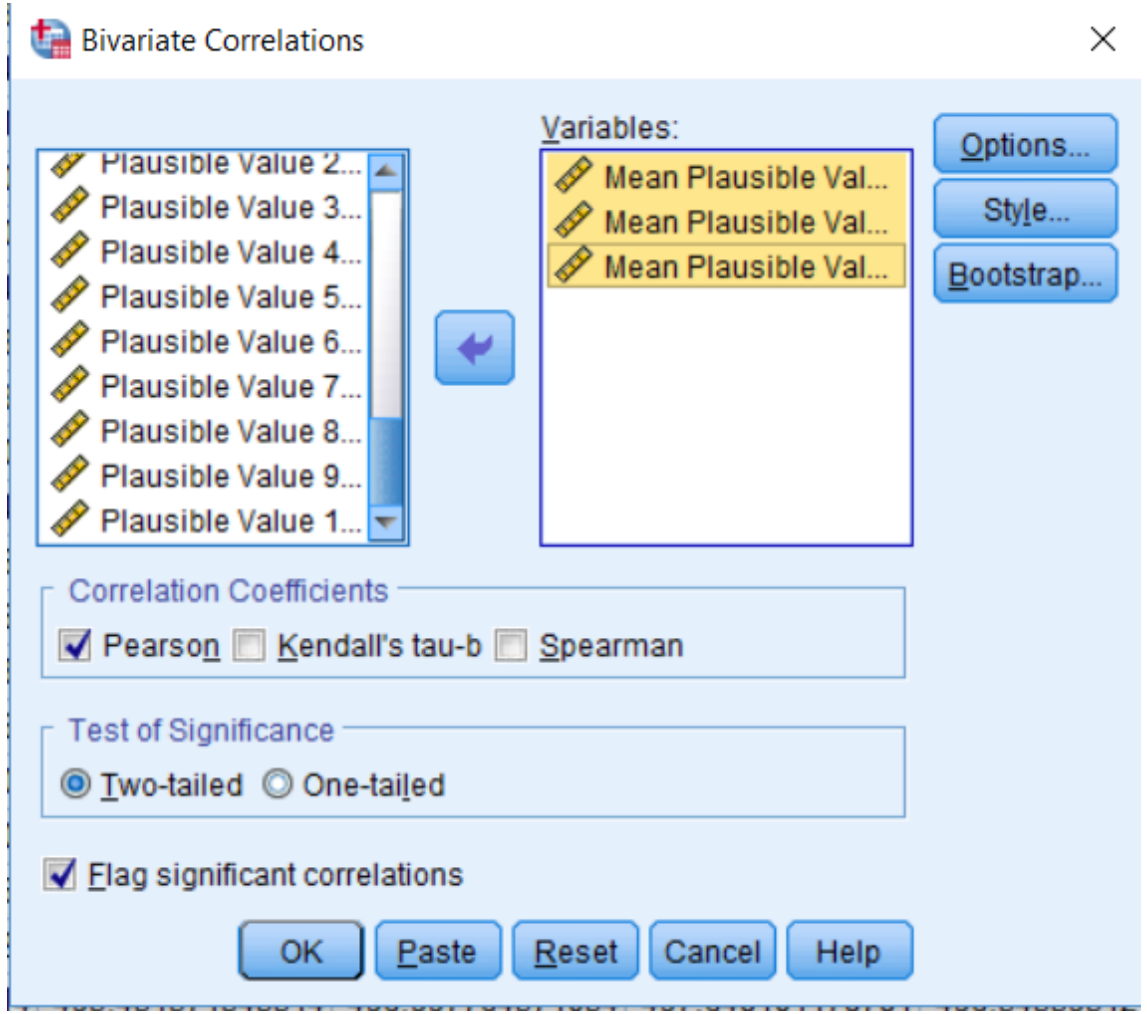


	A	B	C	D	E
1		*Mean Plausible Value in Mathematics*	*Mean Plausible Value in Reading*	*Mean Plausible Value in Science*	
2	Mean Plausible Value in Mathematics	1			
3	Mean Plausible Value in Reading	0,936674148	1		
4	Mean Plausible Value in Science	0,972625261	0,963215102	1	
5					

IBM SPSS Statistics

- Приложение: Статистика, иконометрия
- Начин на работа: Менюта и прозорци; писане на програмен код
- Цена: Платен

- Предимства: Огромен набор от възможни анализи
- Недостатъци: Ниска степен на интерактивност



Correlations

		Mean Plausible Value in Mathematics	Mean Plausible Value in Reading	Mean Plausible Value in Science
Mean Plausible Value in Mathematics	Pearson Correlation	1	,937**	,973**
	Sig. (2-tailed)		,000	,000
	N	73	73	73
Mean Plausible Value in Reading	Pearson Correlation	,937**	1	,963**
	Sig. (2-tailed)	,000		,000
	N	73	73	73
Mean Plausible Value in Science	Pearson Correlation	,973**	,963**	1
	Sig. (2-tailed)	,000	,000	
	N	73	73	73

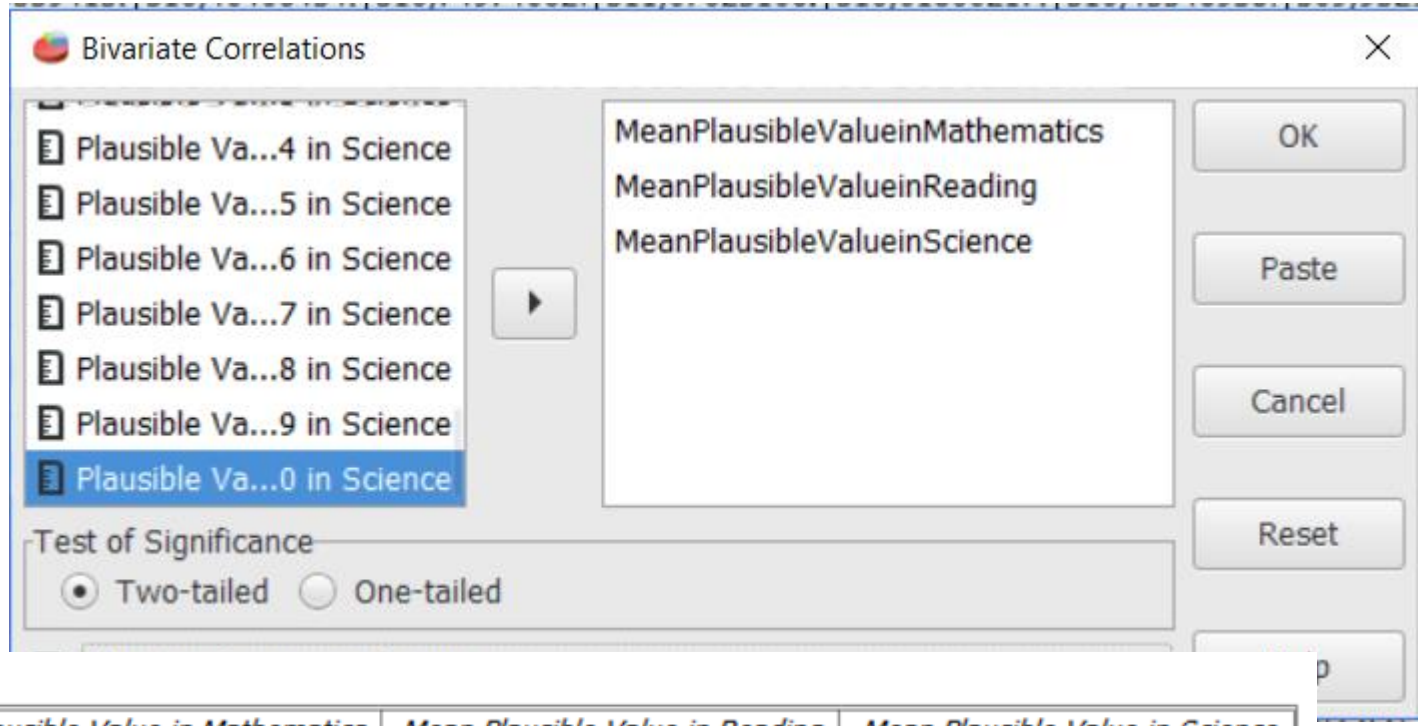
**. Correlation is significant at the 0.01 level (2-tailed).

```
CORRELATIONS
  /VARIABLES=MeanPlausibleValueinMathematics MeanPlausibleValueinReading MeanPlausibleValueinScience
  /PRINT=TWOTAIL NOSIG
  /MISSING=PAIRWISE.
```

PSPP

- Приложение: Статистика
- Начин на работа: Менюта и прозорци
- Цена: Безплатен

- Предимства: Наподобява IBM SPSS Statistics
- Недостатъци: Относително малък набор от възможни анализи; не е интерактивен



Correlations

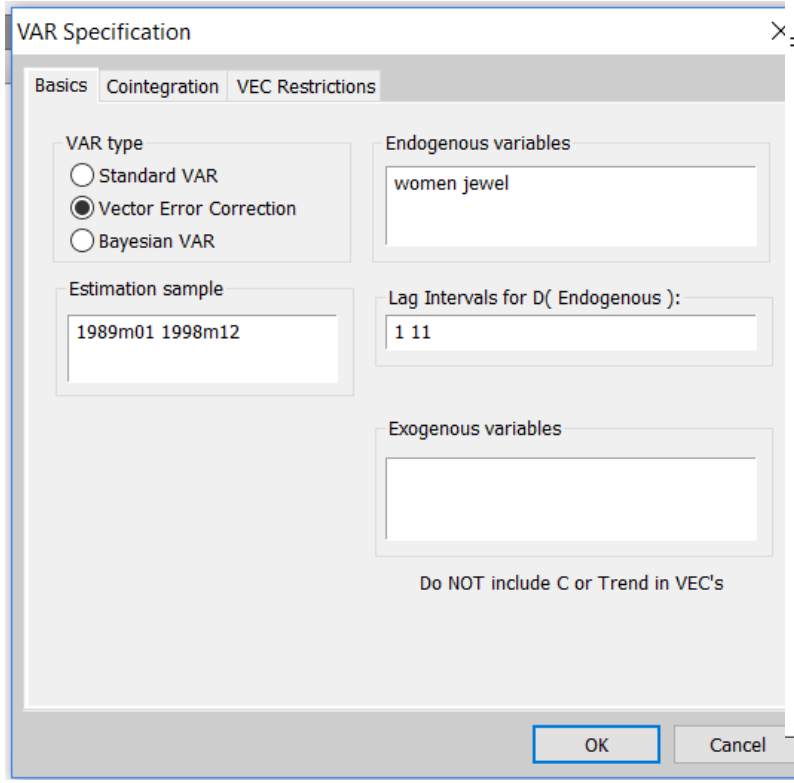
		Mean Plausible Value in Mathematics	*Mean Plausible Value in Reading*	*Mean Plausible Value in Science*
Mean Plausible Value in Mathematics	*Pearson Correlation*	1,00	*,94*	*,97*
	Sig. (2-tailed)		,000	,000
	N	73	73	73
Mean Plausible Value in Reading	*Pearson Correlation*	*,94*	1,00	*,96*
	Sig. (2-tailed)	,000		,000
	N	73	73	73
Mean Plausible Value in Science	*Pearson Correlation*	*,97*	*,96*	1,00
	Sig. (2-tailed)	,000	,000	
	N	73	73	73

EViews

EViews® 10

- Приложение: Иконометрия
- Начин на работа: Менюта и прозорци; писане на програмен код
- Цена: Платен

- Предимства: Ефективни изчисления
- Недостатъци: Проблеми с използването на данни от други продукти



View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Impulse | Resids

Vector Error Correction Estimates
Date: 03/12/18 Time: 22:52
Sample (adjusted): 1990M01 1998M12
Included observations: 108 after adjustments
Standard errors in ( ) & t-statistics in [ ]

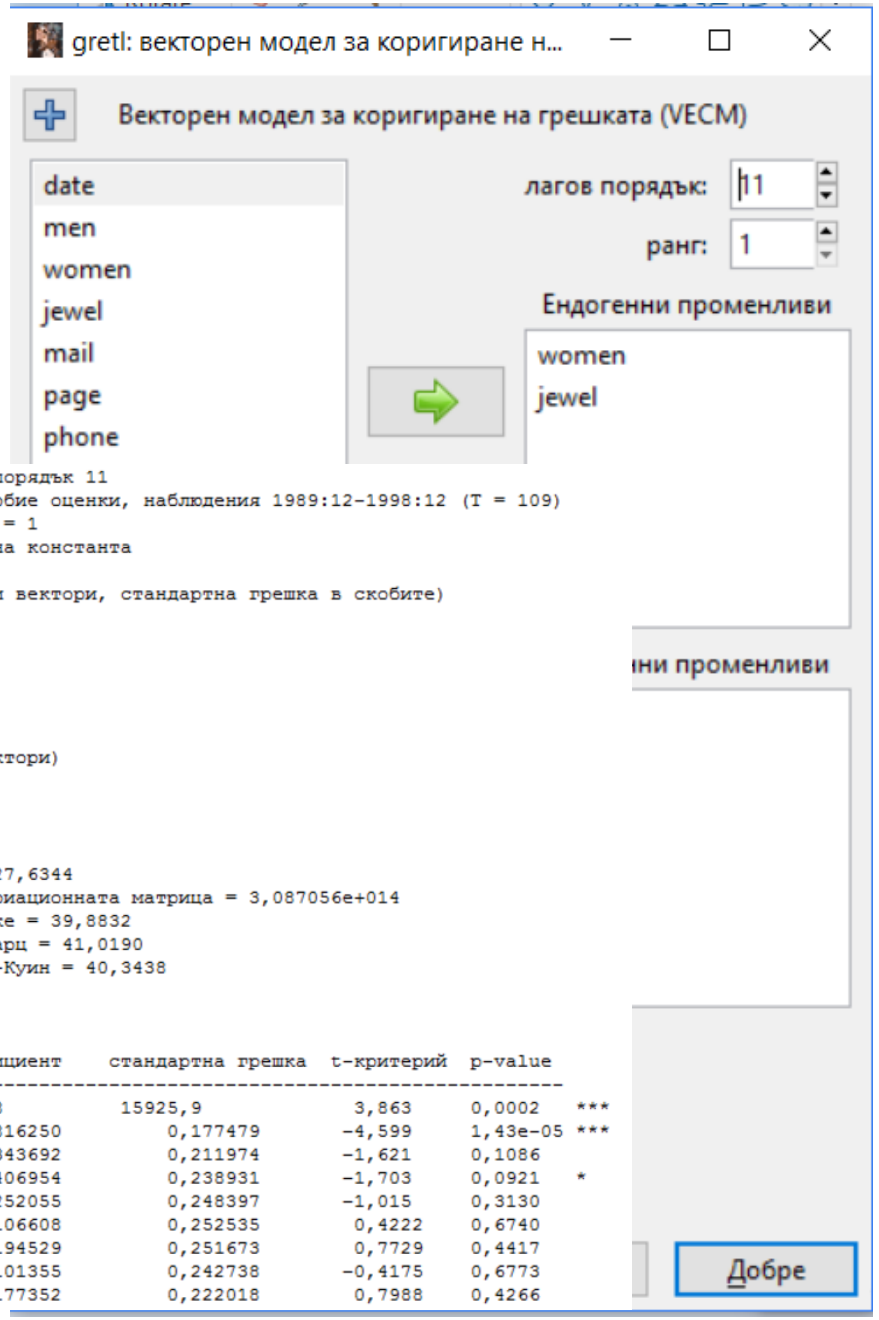
Cointegrating Eq:	CointEq1
WOMEN(-1)	1.000000
JEWEL(-1)	-33.57107 (10.2931) [-3.26153]
C	520459.3

Error Correction:	D(WOMEN)	D(JEWEL)
CointEq1	0.068665 (0.02768) [2.48089]	0.053523 (0.01673) [3.19907]
D(WOMEN(-1))	-0.961166 (0.15470) [-6.21327]	-0.080627 (0.09351) [-0.86220]
D(WOMEN(-2))	-0.802443 (0.19463) [-4.12295]	0.037585 (0.11765) [0.31946]
D(WOMEN(-3))	-0.805171 (0.21540) [-3.73799]	-0.104838 (0.13021) [-0.80515]
D(WOMEN(-4))	-0.628510 (0.22110) [-2.84259]	-0.008221 (0.13366) [-0.06151]
D(WOMEN(-5))	-0.298444 (0.22719) [-1.31365]	0.147138 (0.13733) [1.07139]

Gretl

- Приложение: Иконометрия
- Начин на работа: Менюта и прозорци; писане на програмен код
- Цена: Безплатен

- Предимства: Наподобява EViews
- Недостатъци: Ограничения за обема на данните



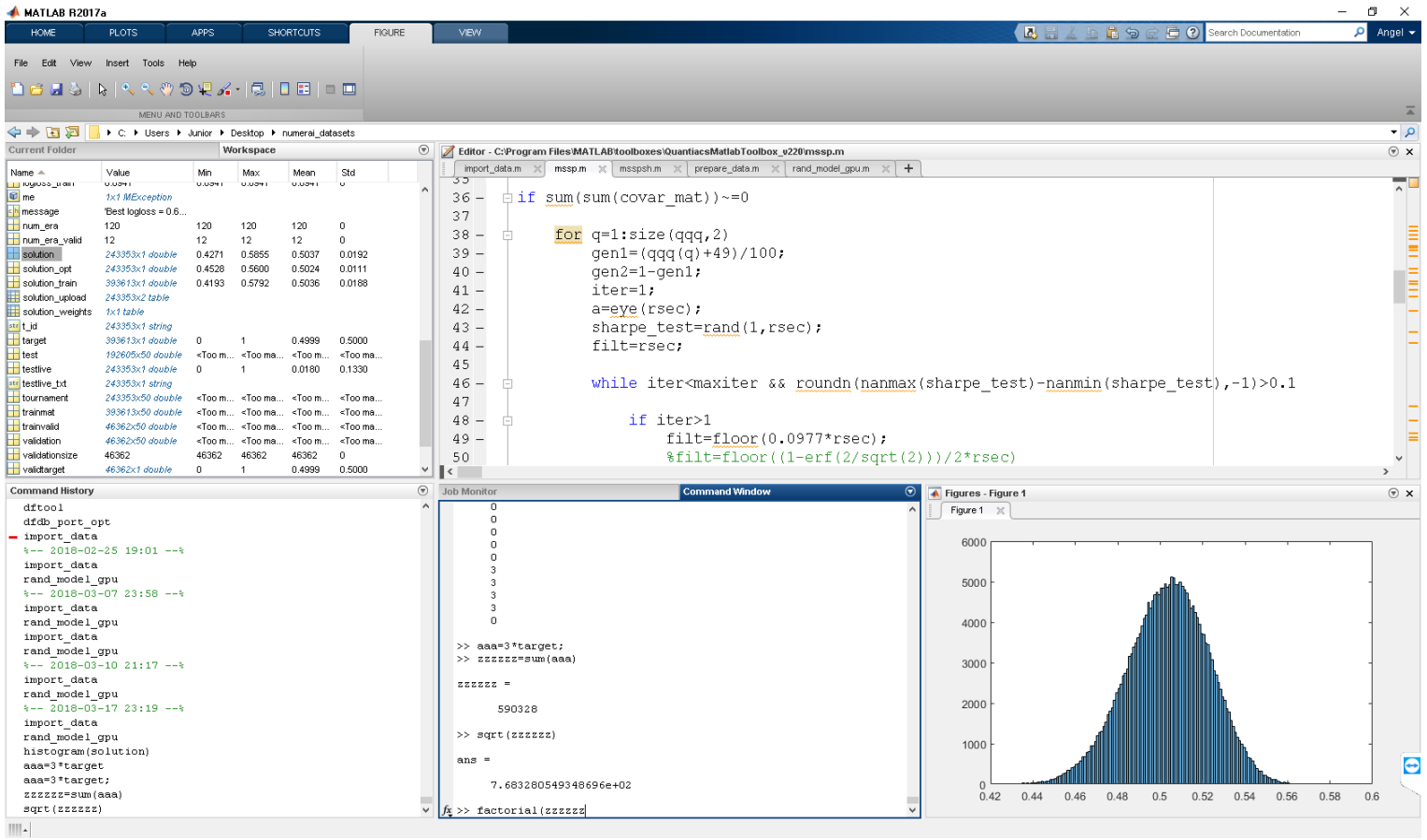
The All-Stars

MatLab

MATLAB

- Приложение: Статистика; иконометрия
- Начин на работа: Писане на програмен код
- Цена: Платен

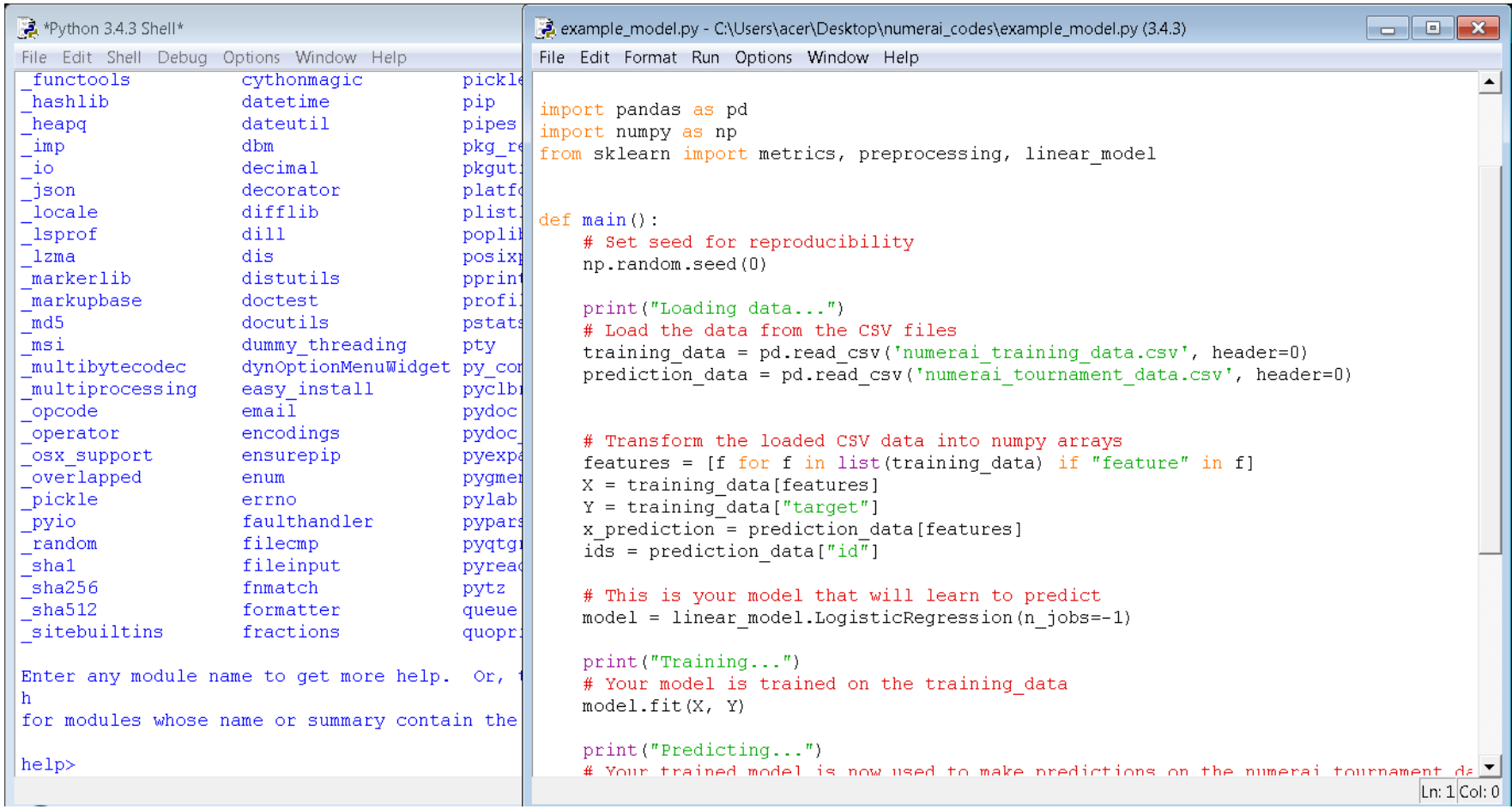
- Предимства: Много добра документация; паралелни изчисления
- Недостатъци: Много висока цена



Python

- Приложение: Статистика; иконометрия; data mining
- Начин на работа: Писане на програмен код
- Цена: Безплатен

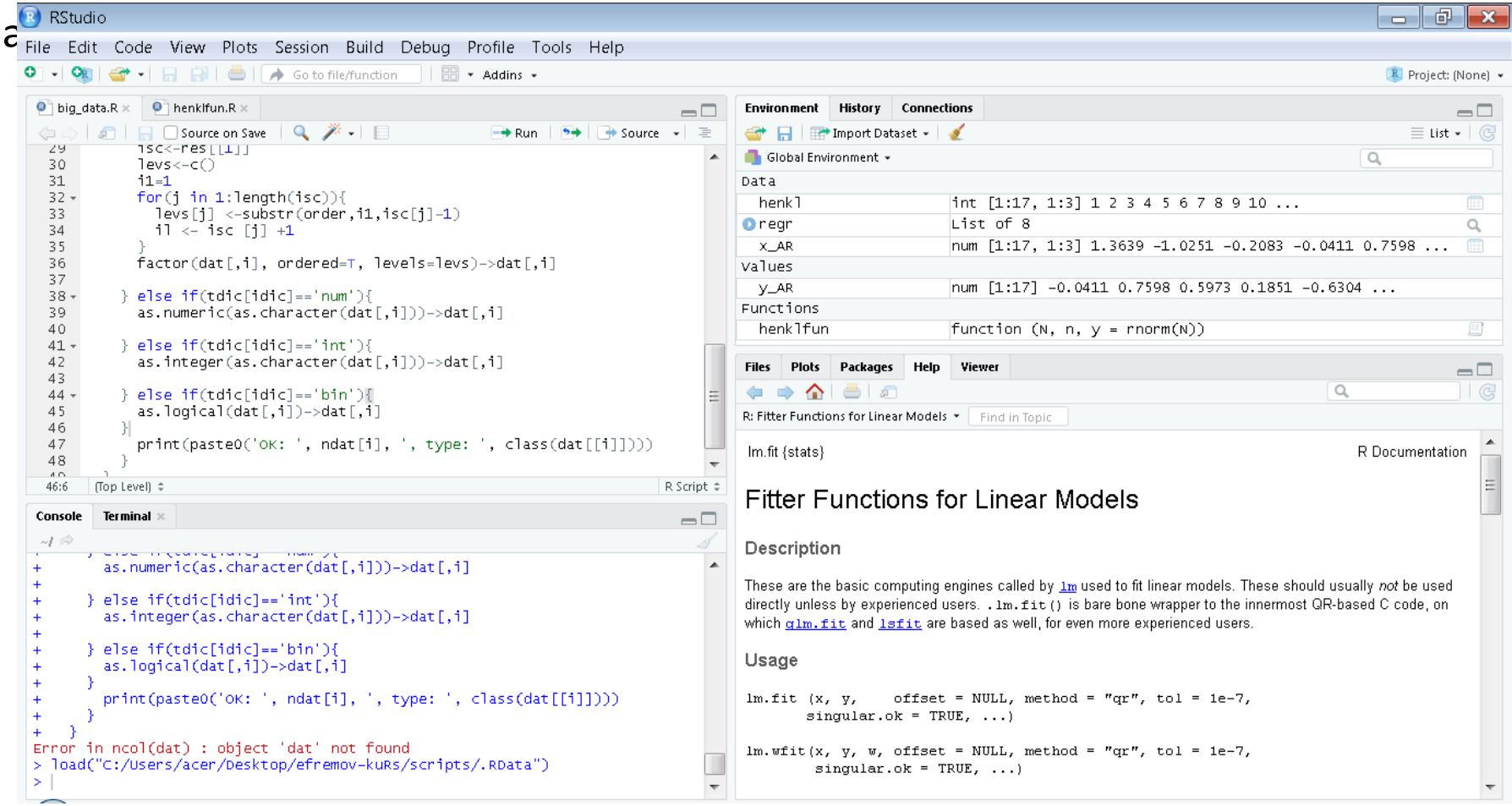
- Предимства: Глобална общност от разработчици
- Недостатъци:



R (+R studio)

- Приложение: Статистика; иконометрия; data mining
- Начин на работа: Писане на програмен код
- Цена: Безплатен

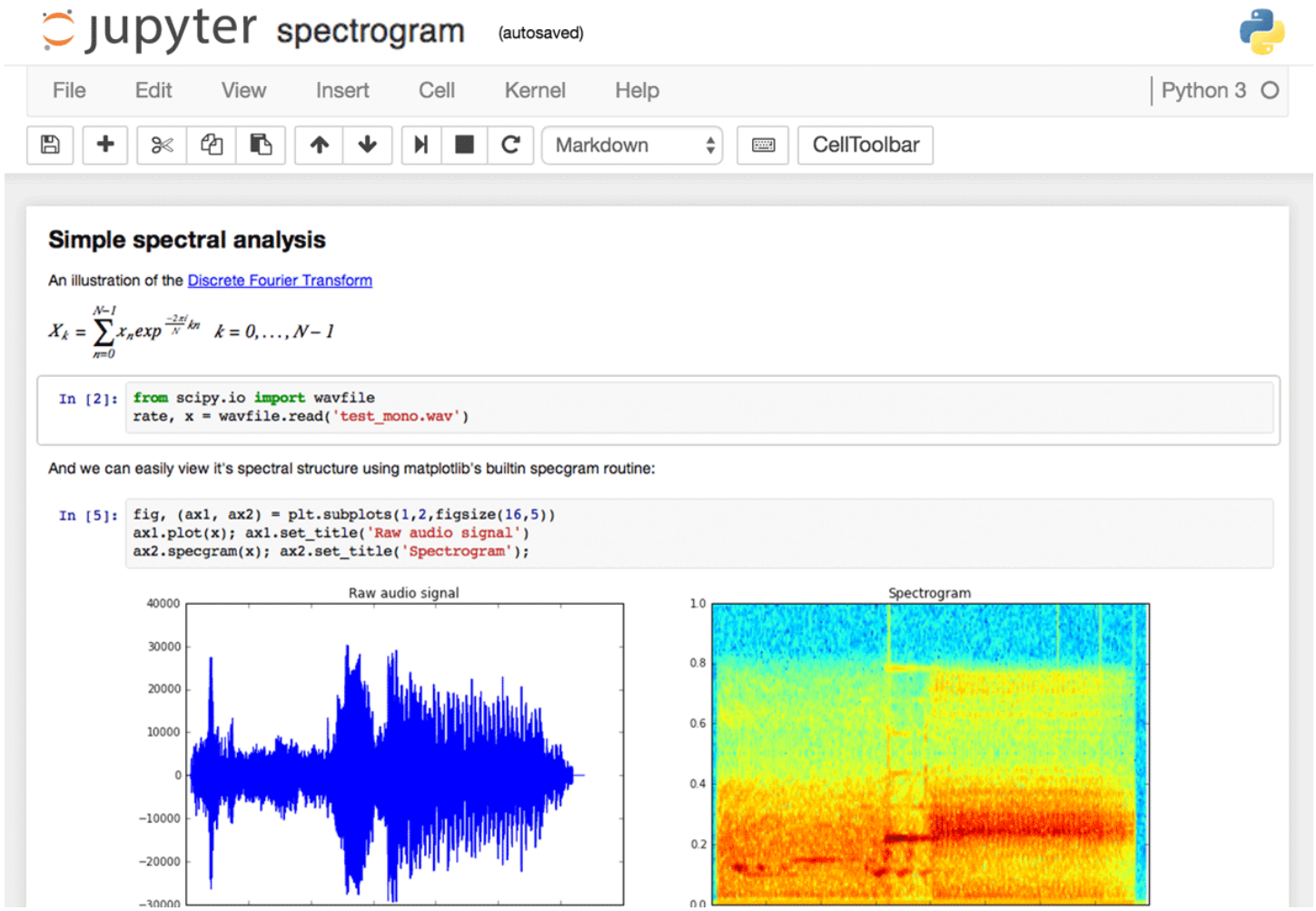
- Предимства: Глобална общност от разработчици
- Недостатъци: Малко странен език



Jupyter Notebook

- Приложение: Data mining
- Начин на работа: Онлайн
- Цена: Безплатен

- Предимства: Индустриален стандарт за Data Science
- Недостатъци:

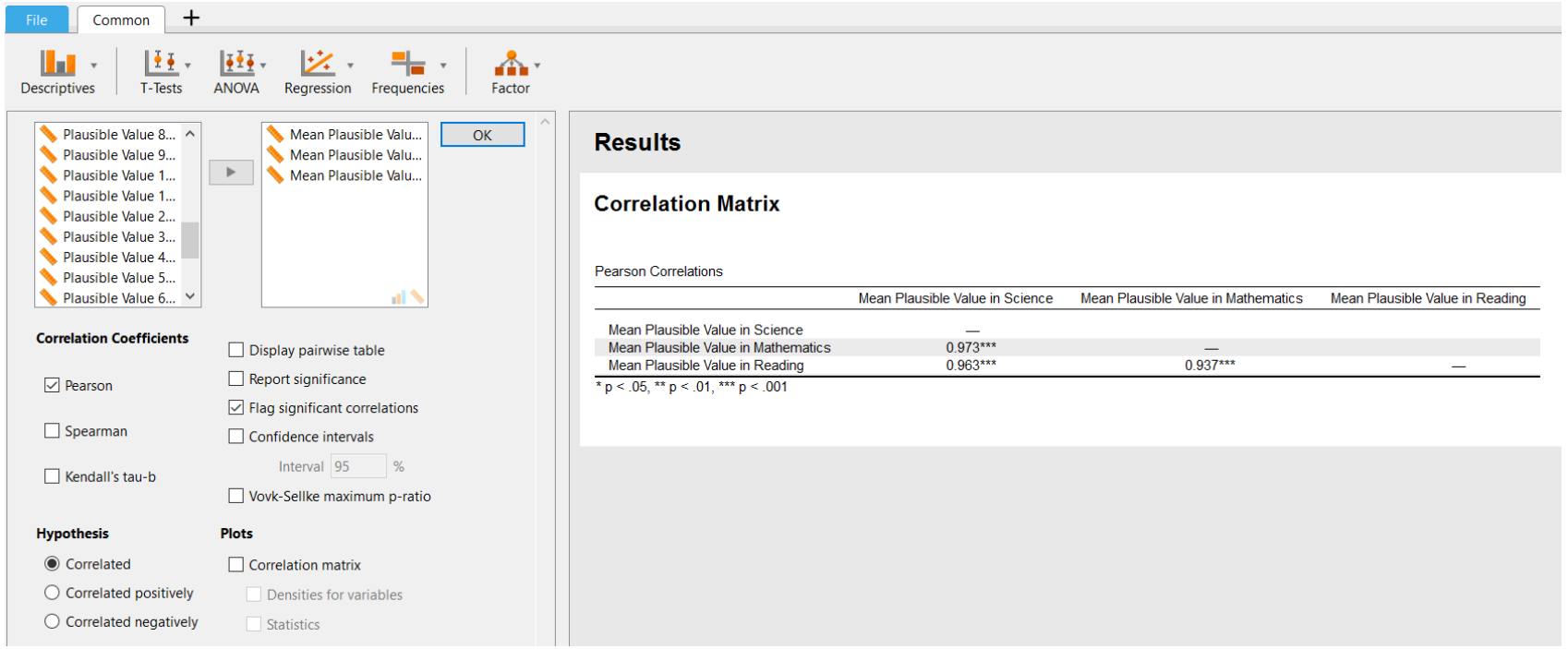


The User-Friendlies

JASP

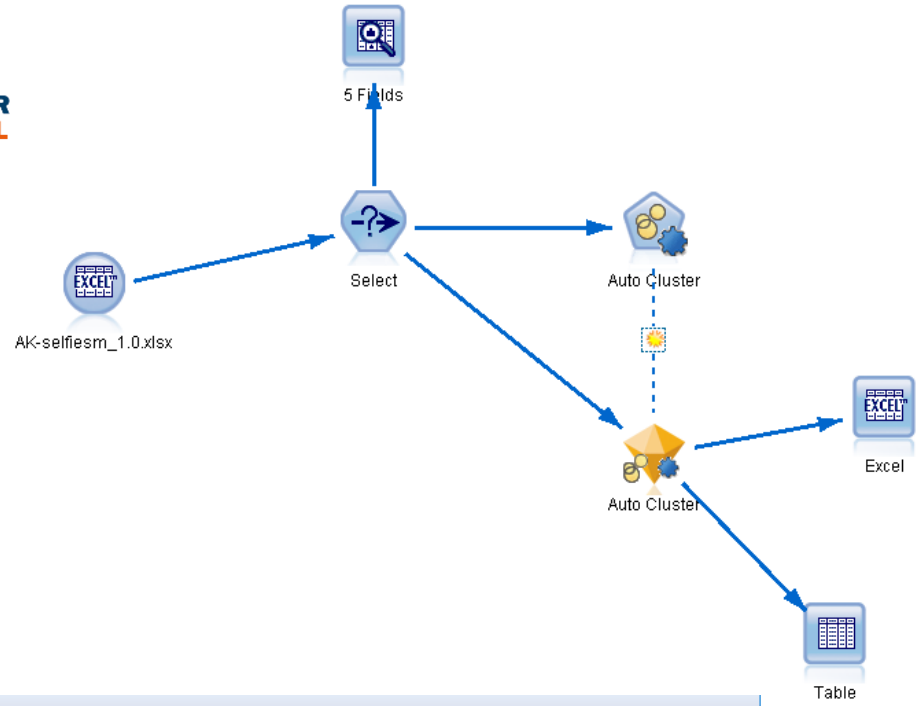
- Приложение: Статистика
- Начин на работа: Менюта и прозорци
- Цена: Безплатен

- Предимства: Висока степен на интерактивност
- Недостатъци: Относително малък набор от възможни анализи



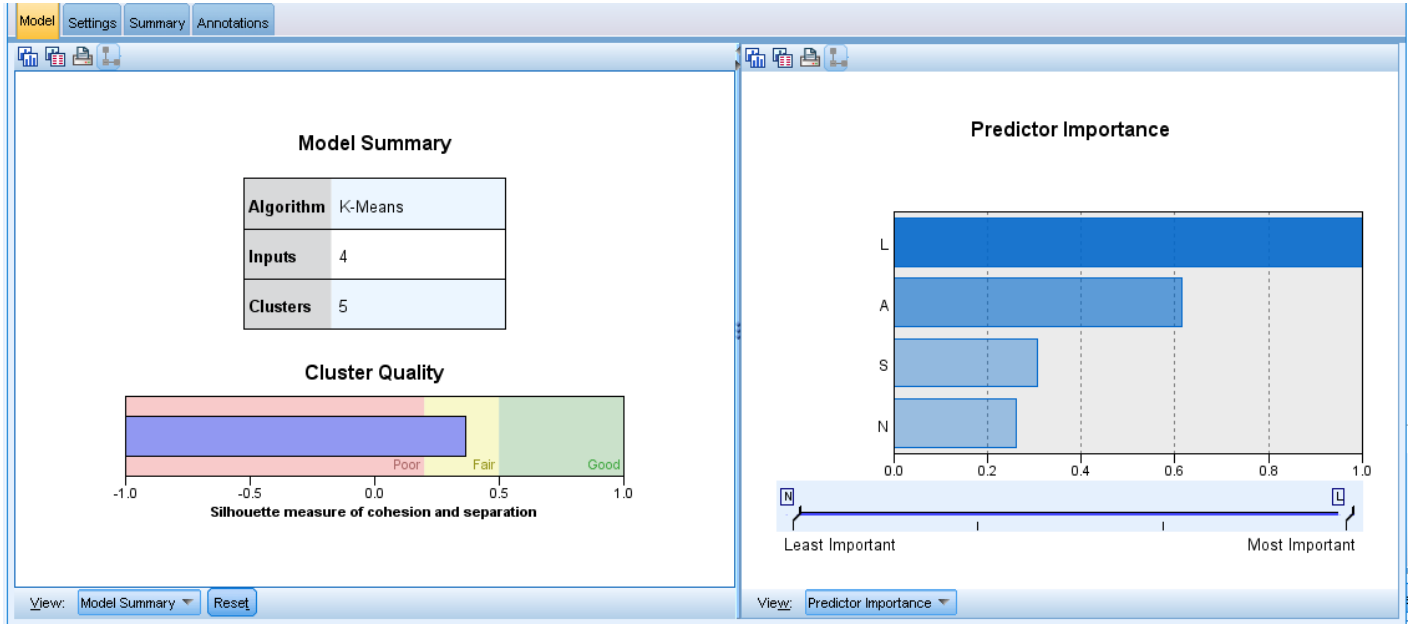
IBM SPSS Modeler

IBM SPSS MODELER PROFESSIONAL



- Приложение: Иконометрия; data mining
- Начин на работа: Графични обекти
- Цена: Платен

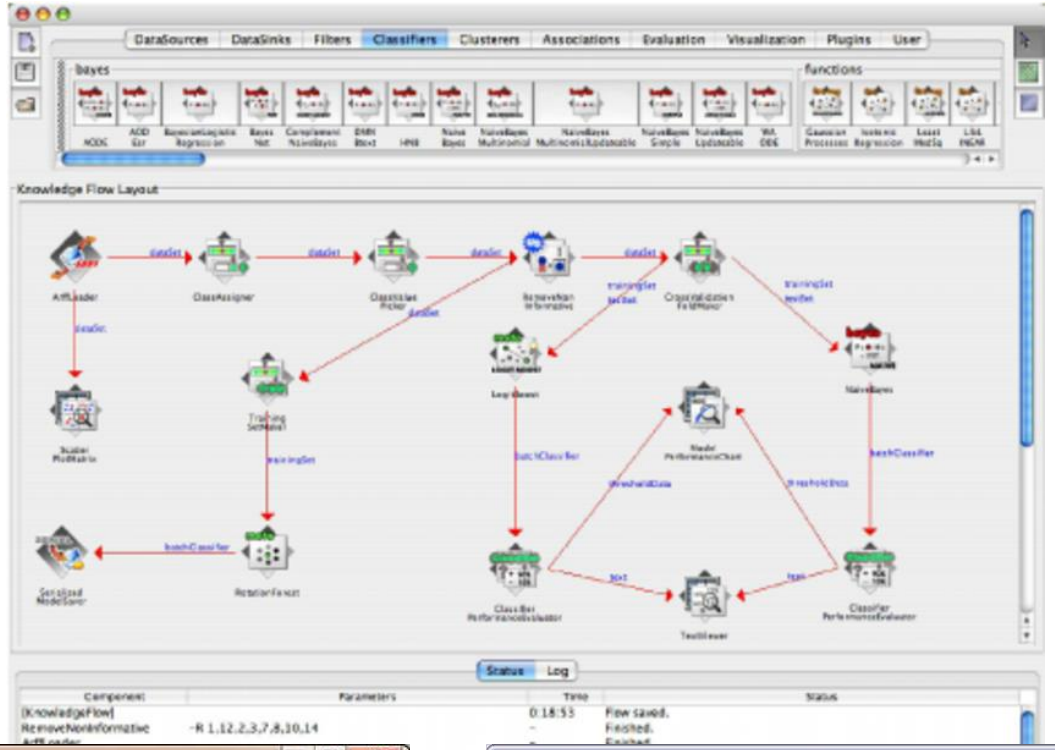
- Предимства: Добро използване на ресурсите
- Недостатъци: Не е много user friendly, когато се работи с много променливи

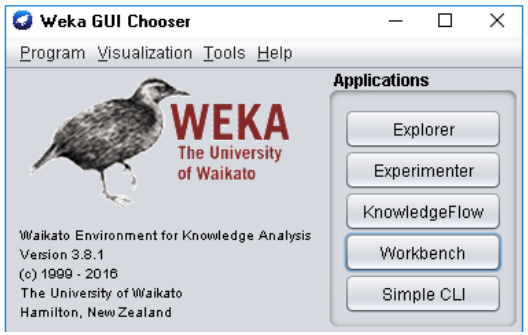


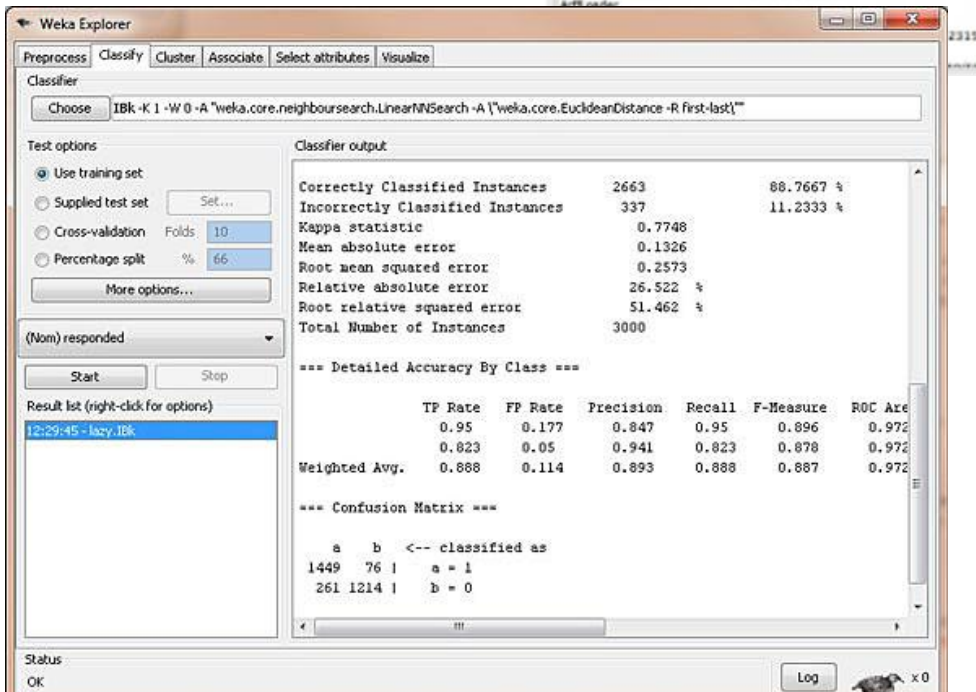
Weka

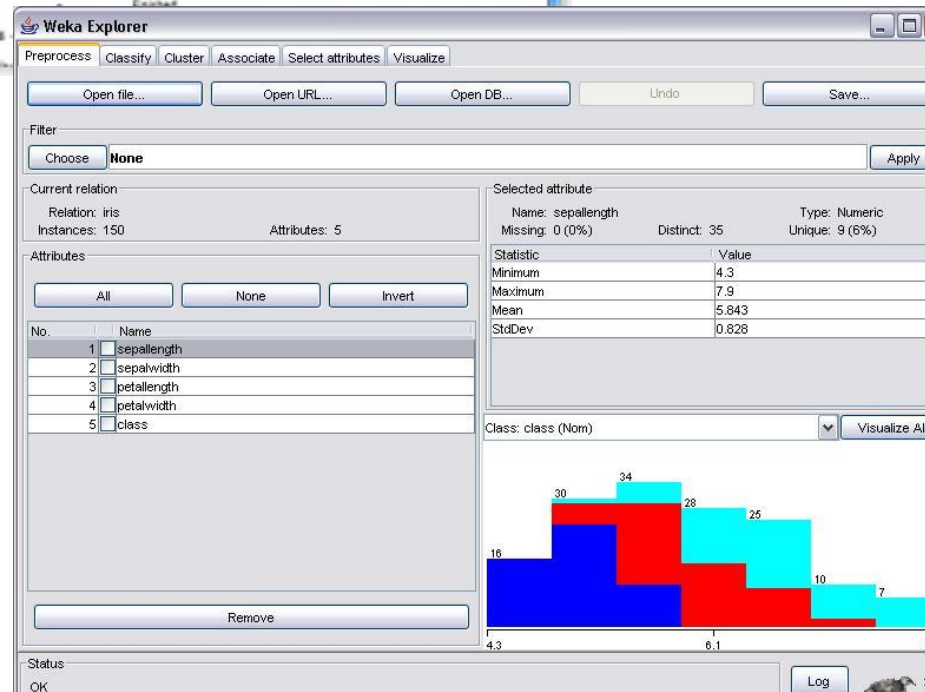
- Приложение: Статистика; data mining
- Начин на работа: Графични обекти
- Цена: Безплатен

- Предимства: Един от пионерите в областта
- Недостатъци: Демодиран и тромав





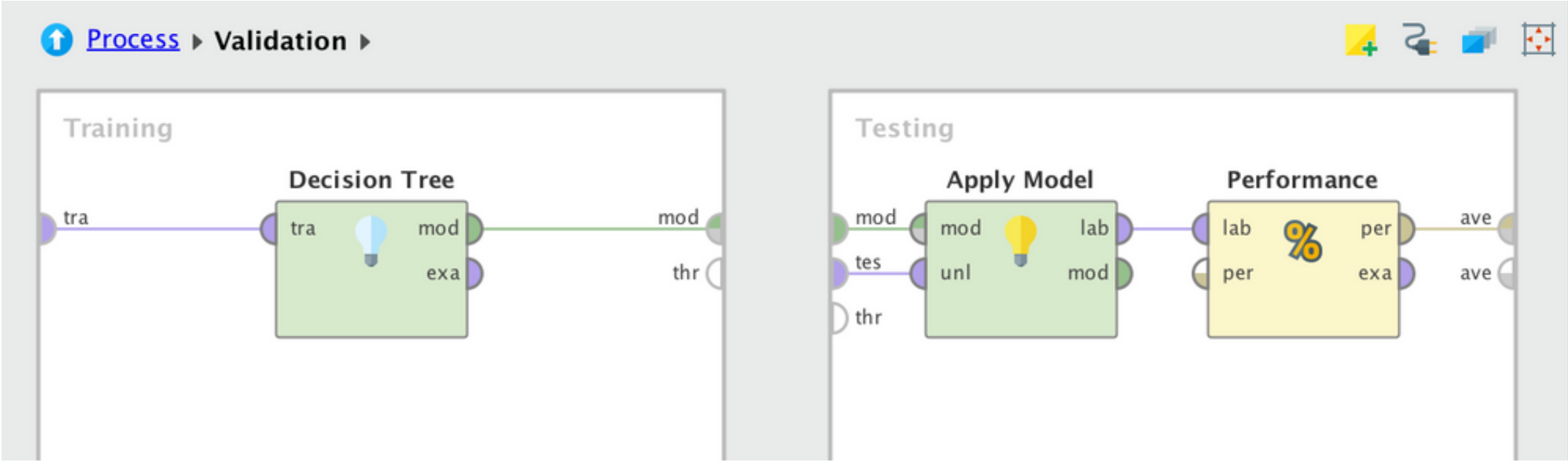




Rapid Miner

- Приложение: Статистика; data mining
- Начин на работа: Графични обекти
- Цена: Платен

- Предимства: Вероятно най-интуитивният интерфейс
- Недостатъци:

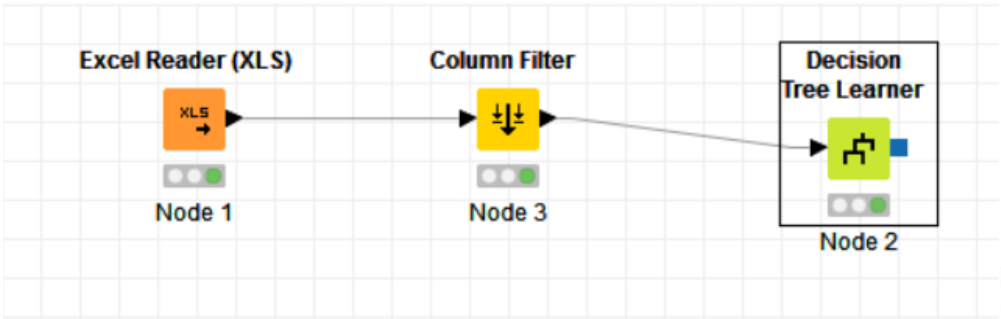


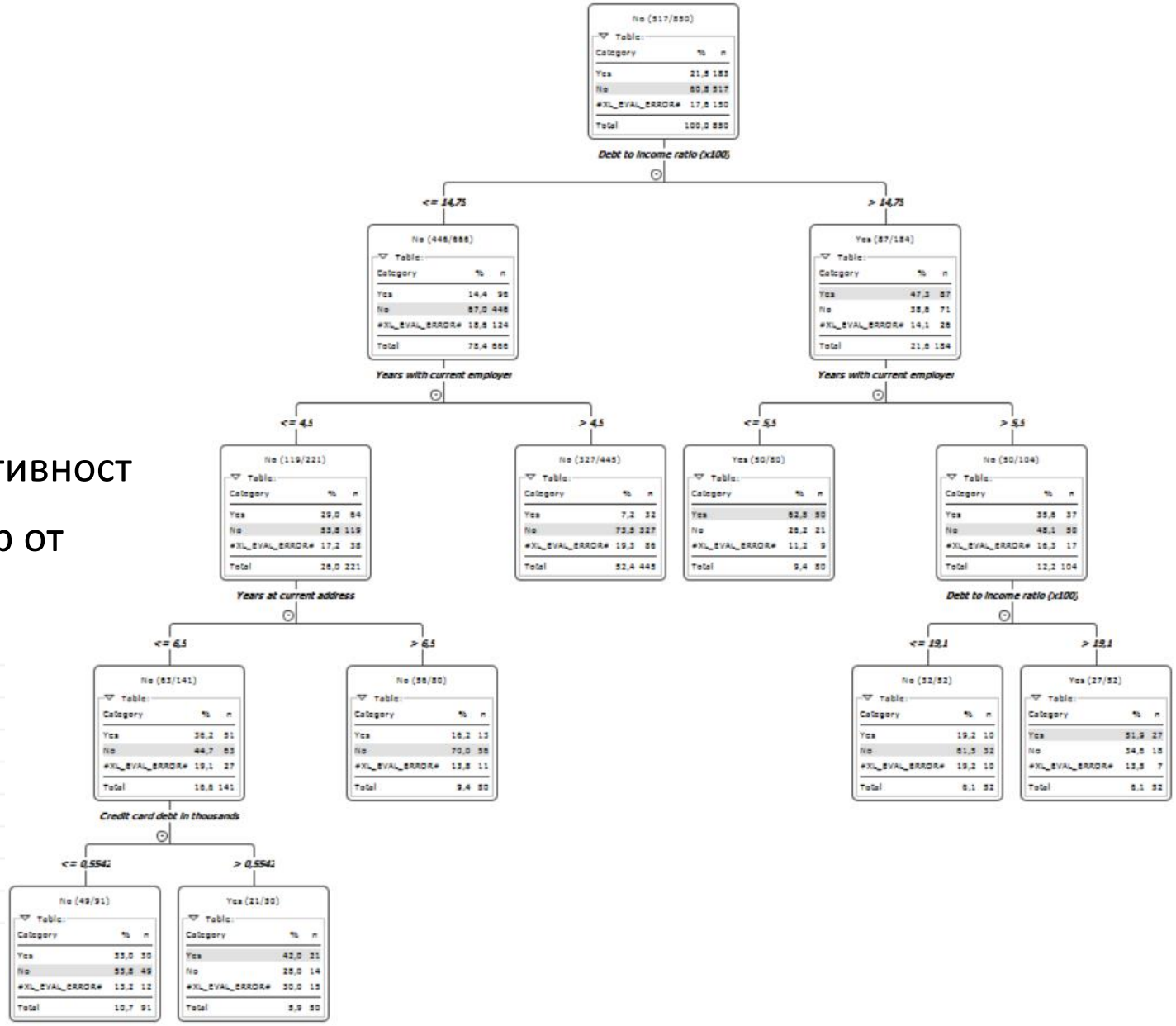
KNIME

KNIME

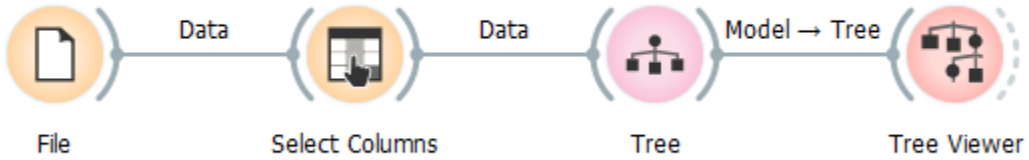
- Приложение: Статистика; data mining
- Начин на работа: Графични обекти
- Цена: Безплатен

- Предимства: Висока степен на интерактивност
- Недостатъци: Относително малък набор от възможни анализи



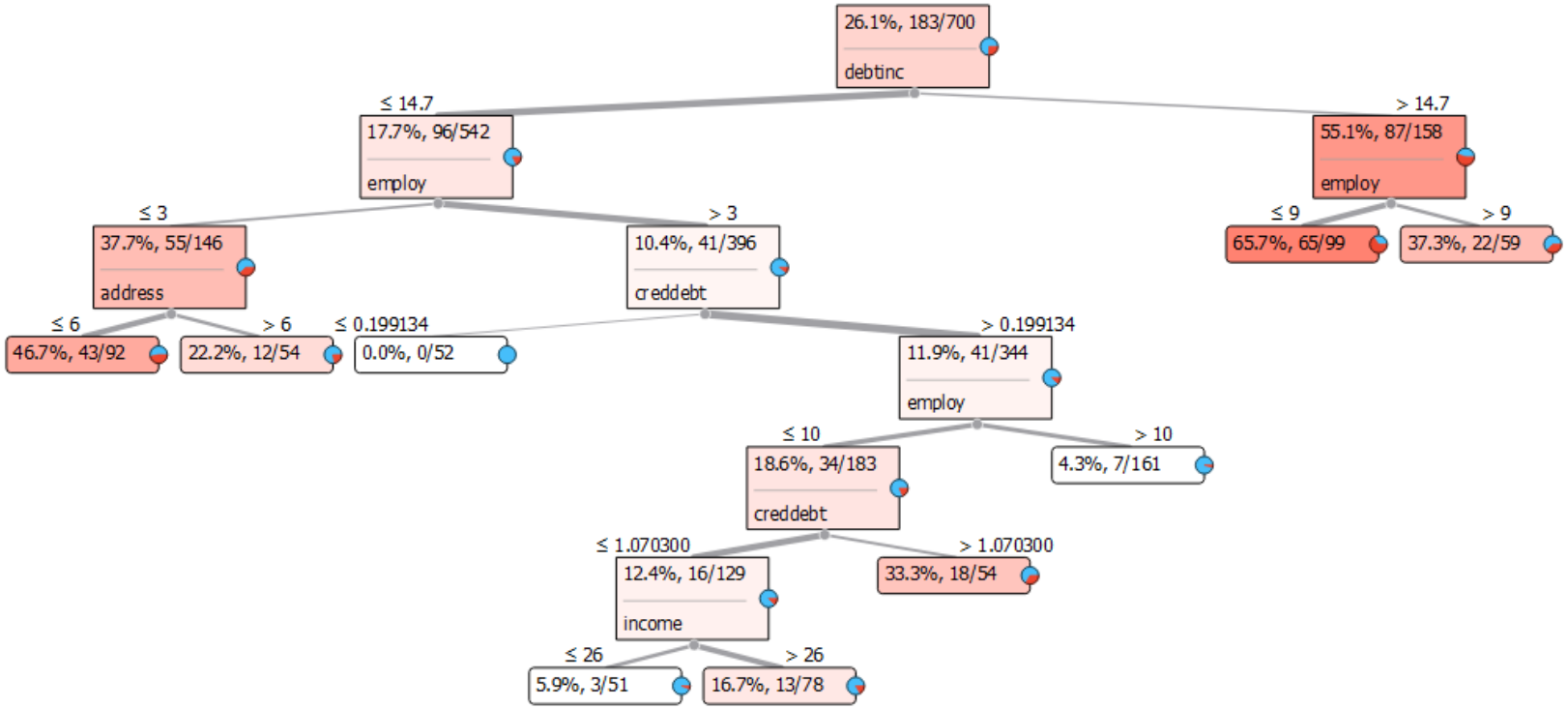


Orange



- Приложение: data mining
- Начин на работа: Графични обекти
- Цена: Безплатен

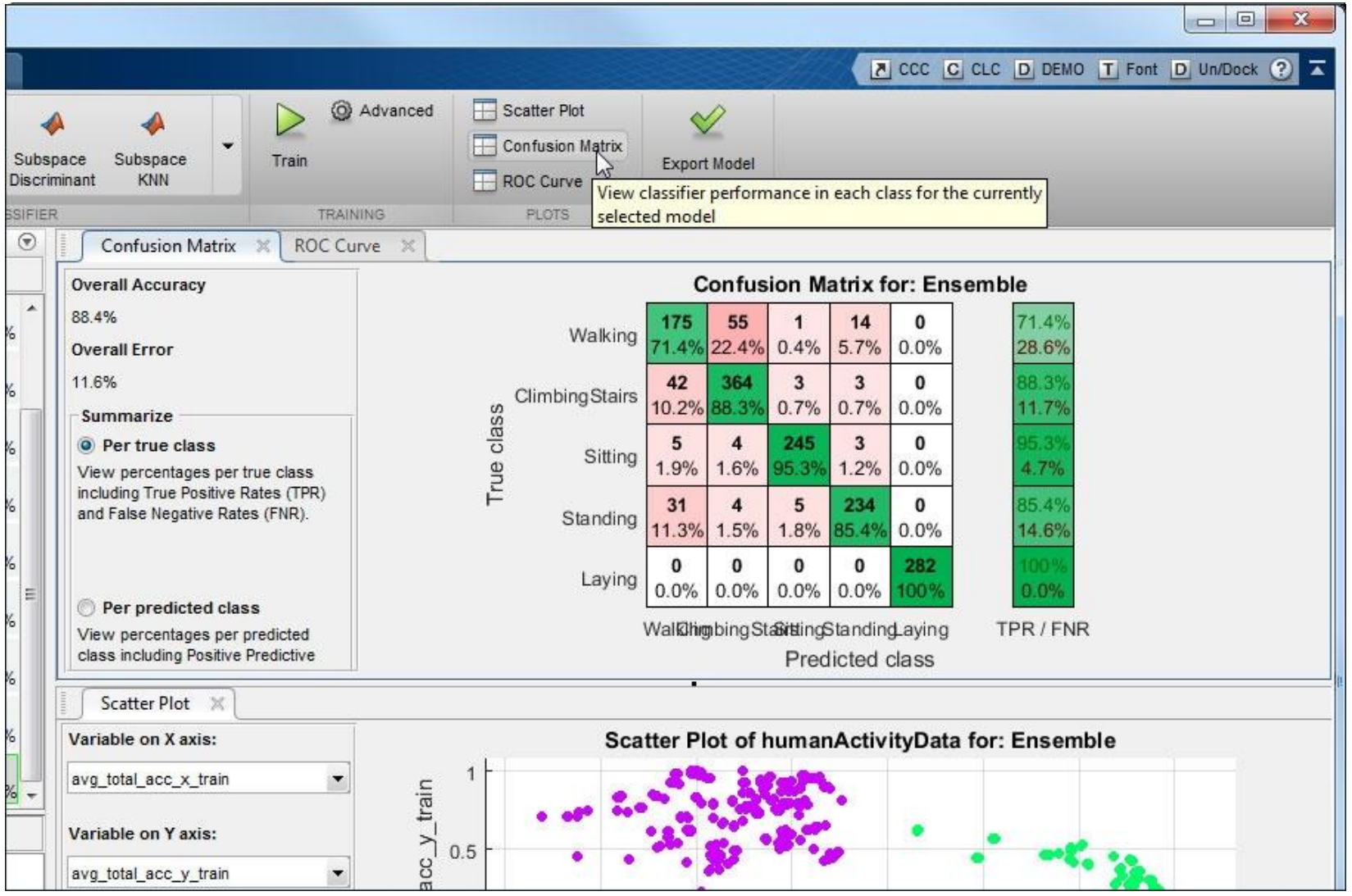
- Предимства: Висока степен на интерактивност
- Недостатъци: Относително малък набор от възможни анализи



MatLab Classification Learner

- Приложение: data mining
- Начин на работа: Графични обекти
- Цена: Платен

- Предимства: Част от средата на Matlab
- Недостатъци: Все още се разработва, за да включи още модели



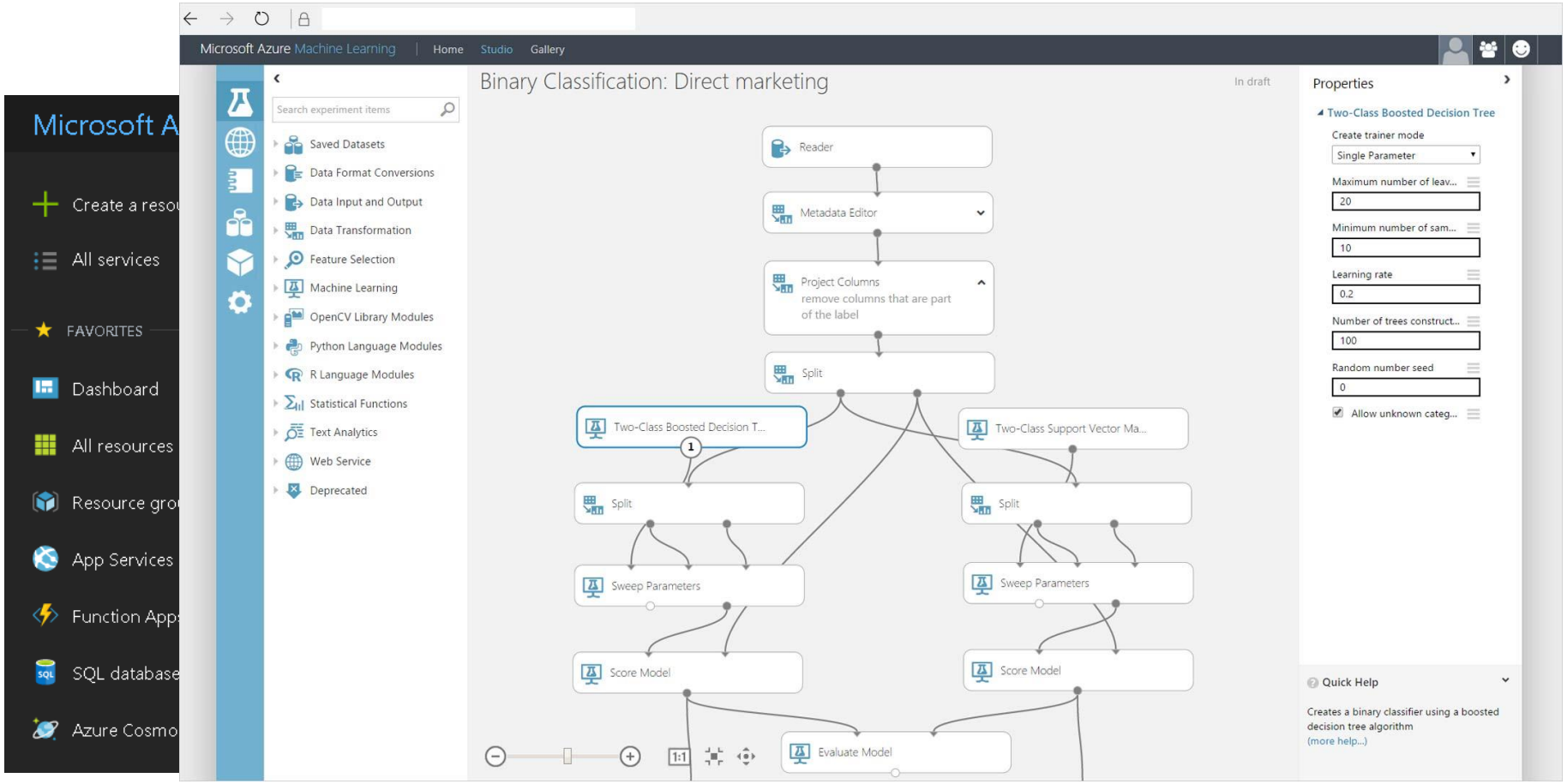
The On-liners

Microsoft Azure

Azure

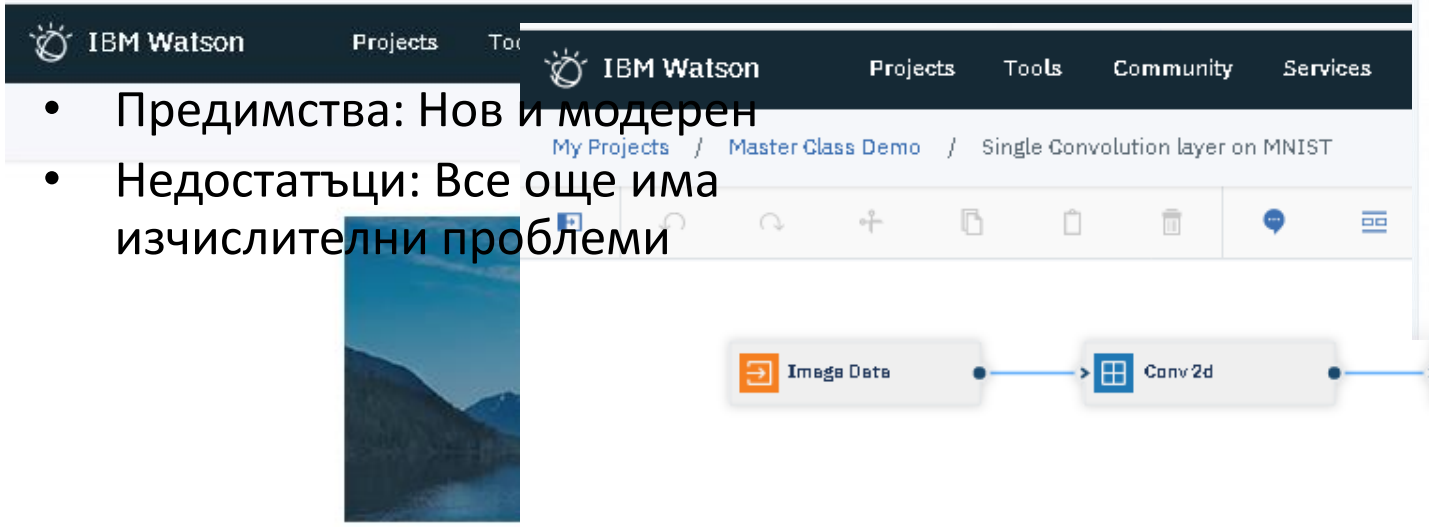
- Приложение: data mining
- Начин на работа: онлайн
- Цена: Платен

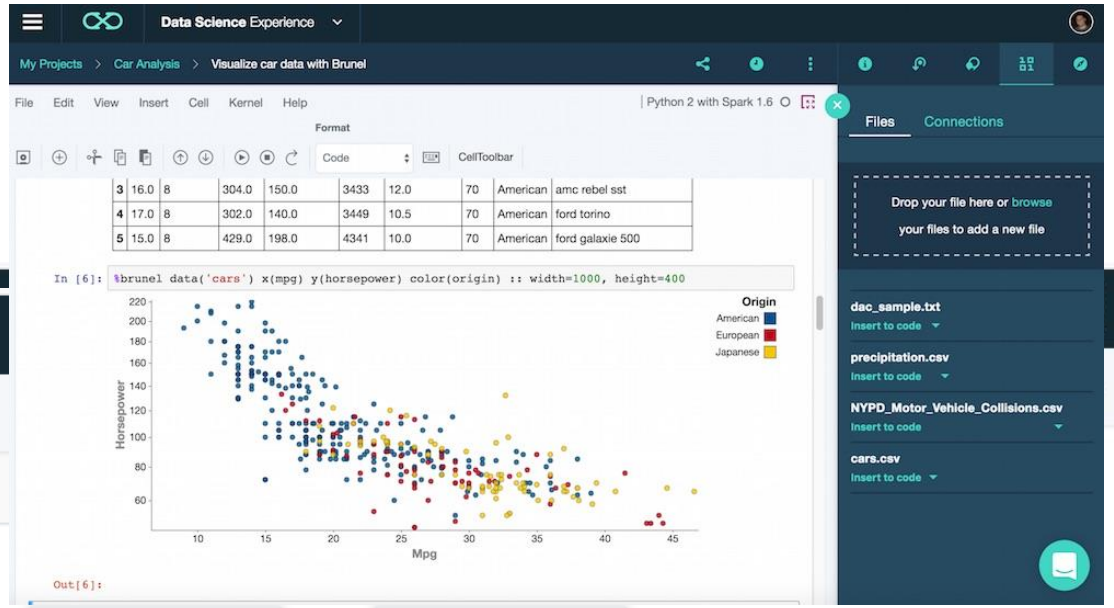
- Предимства: Вече са достъпни много инструменти
- Недостатъци: Може би е малко труден за инсталиране

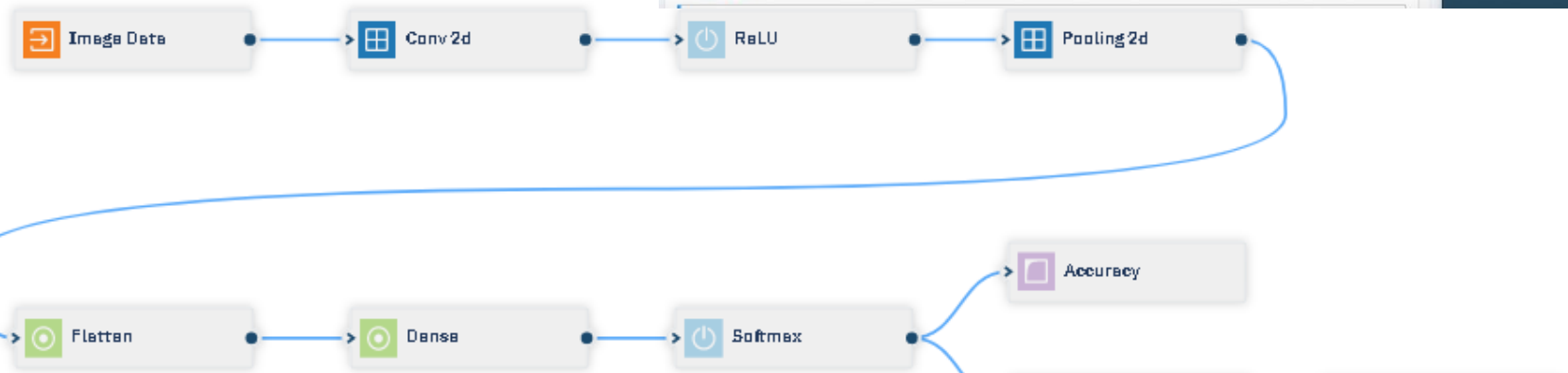


IBM Watson Studio

- Приложение: data mining
- Начин на работа: онлайн
- Цена: Платен
- Предимства: Нов и модерен
- Недостатъци: Все още има изчислителни проблеми







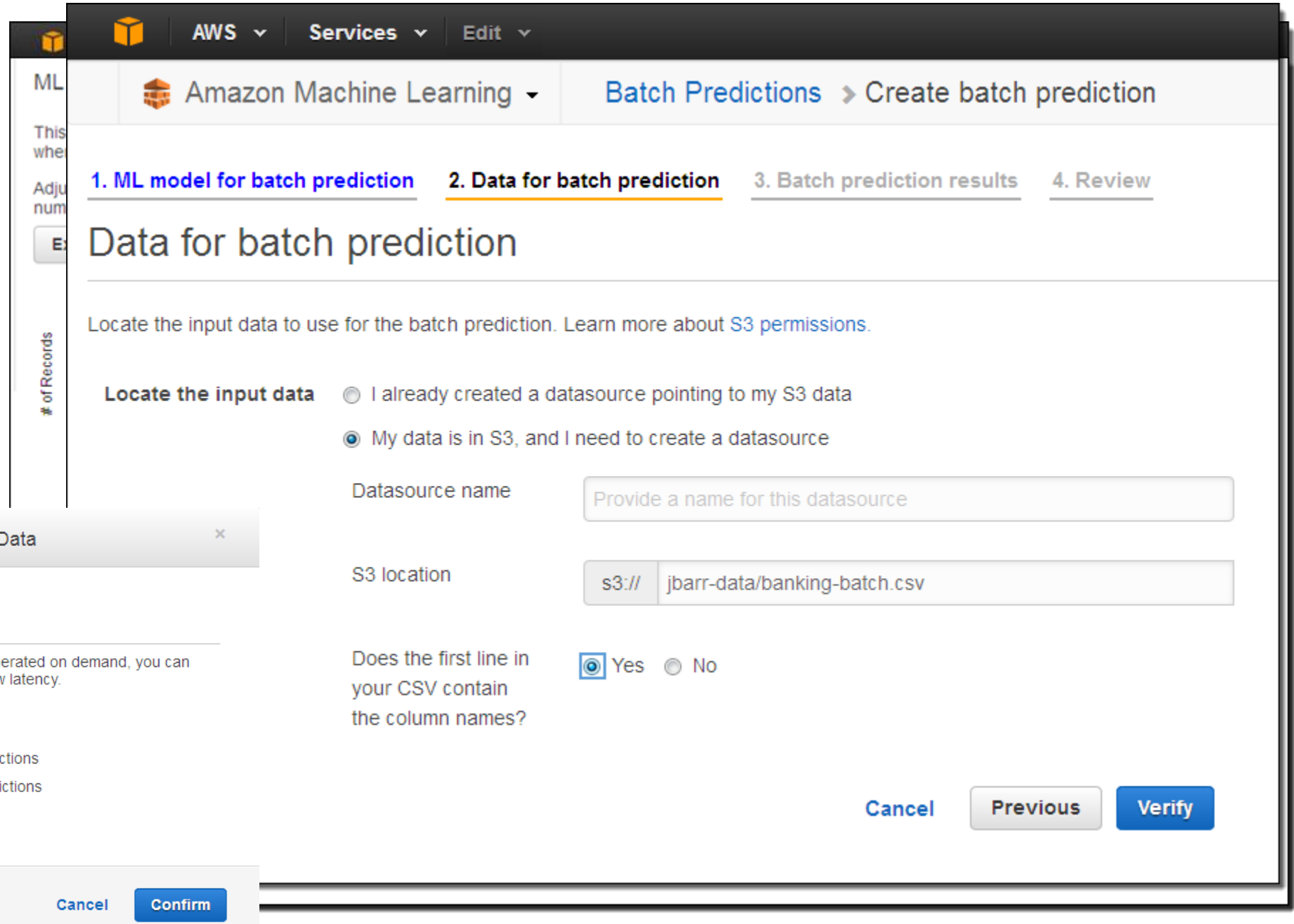
Get started with

Refine data | New notebook | Deep learning | New Modeler flow | New model | New streams flow

Amazon ML

- Приложение: data mining
- Начин на работа: онлайн
- Цена: Платен

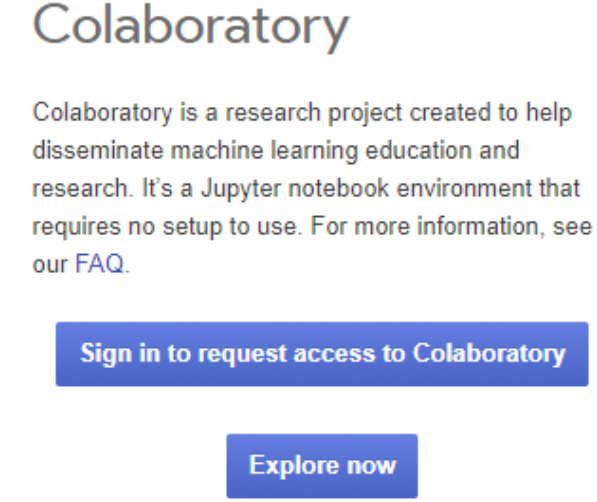
- Предимства: Интегриран с AWS S3 и може да работи в реално време
- Недостатъци: Все още се разработва, за да включи още модели

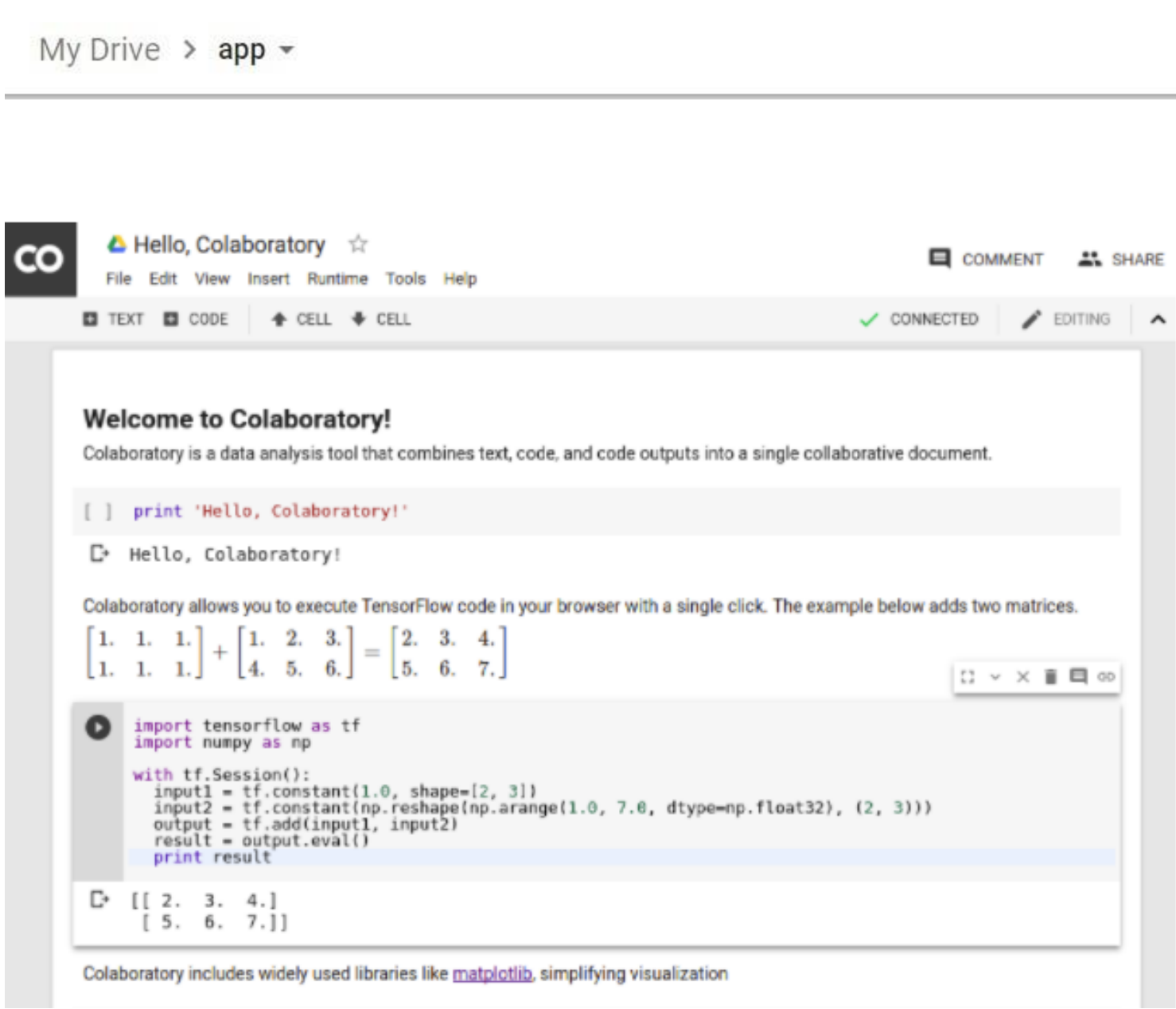


Google Colab

- Приложение: data mining
- Начин на работа: онлайн
- Цена: Безплатен

- Предимства: GPU изчисления чрез Tensor Flow
- Недостатъци: 12-часови сесии



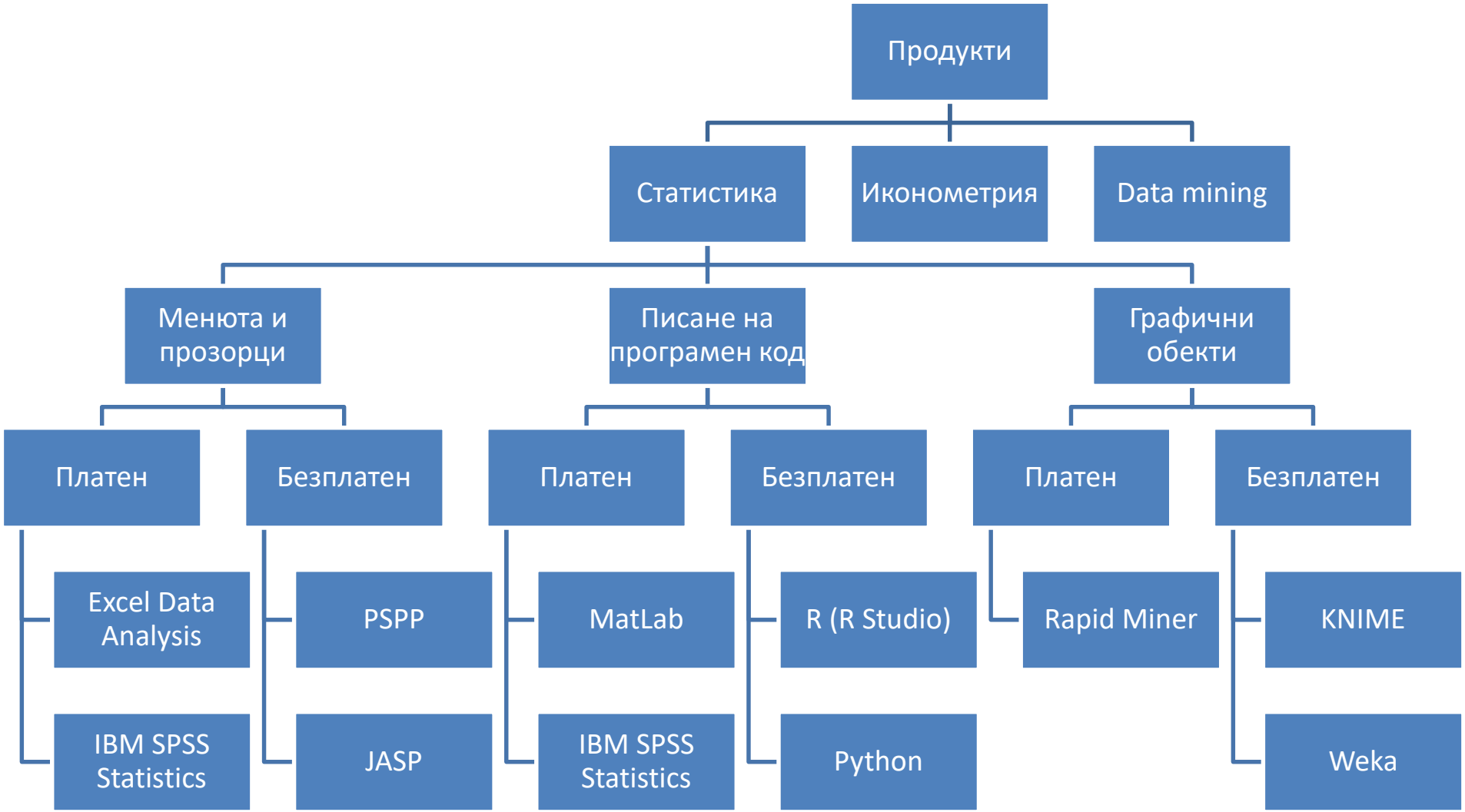


Критерии за избор на точния продукт

- Какъв проблем решавате? (Приложение)
- Как искате да работите? (Начин на работа)
- Колко сте склонни да платите? (Цена)

Приложение	Начин на работа	Цена	Продукт		
Статистика	Менюта и прозорци	Платен	Excel Data Analysis	IBM SPSS Statistics	
		Безплатен	PSPP	JASP	
	Писане на програмен код	Платен	MatLab	IBM SPSS Statistics	
		Безплатен	R (+ R Studio)	Python	
	Графични обекти	Платен	Rapid Miner		
		Безплатен	KNIME	Weka	
Иконометрия	Менюта и прозорци	Платен	eViews	IBM SPSS Statistics	
		Безплатен	Gretl		
	Писане на програмен код	Платен	eViews	IBM SPSS Statistics	MatLab
		Безплатен	Gretl	R (+ R Studio)	Python
	Графични обекти	Платен	IBM SPSS Modeler		
Data mining	Писане на програмен код	Платен	Matlab		
		Безплатен	R (R Studio)	Python	
	Графични обекти	Платен	IBM SPSS Modeler	Rapid Miner	Matlab Classification App
		Безплатен	Orange	KNIME	Weka
	Онлайн	Платен	IBM Watson Studio	Microsoft Azure	Amazon ML
		Безплатен	Google Colab	Jupyter Notebook	

Дърво на решенията



Дърво на решенията