

Use of R in outliers detection

Audrius Taraila
audrius.taraila@stat.gov.lt

Statistics Lithuania

- ▶ “Shiny” is one of the numerous R packages. This package makes it easy to build interactive web apps straight from R.

Shiny structure

- ▶ Consists of two components
 - User interface object (ui.R file)
 - Server function (server.R file)

ui.R file example

```
1 shinyUI(fluidPage(  
2   tags$head(tags$style(HTML('.skin-blue, .box-title { font-family: "Geneva"!important;}  
3     .content-wrapper,  
4     .right-side {background-color: #ffffff !important;}  
5     '))),  
6   dashboardPage(skin = "blue",  
7     dashboardHeader(title = "Shiny aplikacija duomenÅ³ redagavimui"),  
8     dashboardSidebar(  
9       sidebarMenu(  
10        menuItem("DuomenÅ³ failo Å³kÅ³limas", tabName = "Å³kÅ³limas", icon = icon("table")),  
11        menuItem("PagrindinÅ³s charakteristikos", tabName = "char_summary", icon = icon("bar-ch-  
12        menuItem("IA³skirÅ³iÅ³ radimas", tabName = "isskirtys", icon = icon("wrench"))  
13      )  
14    ),  
15    dashboardBody(  
16      useShinyjs(),  
17      tabItems(  
18        tabItem(tabName = "ikelimas",  
19          # fluidRow(  
20            #   box(width = 4,  
21              #     actionButton("ikelimoBoxButton", "Rodyti/PaslÅ³pti Å³kÅ³limo skiltÅ³")  
22            #   )  
23          # ),  
24          fluidRow(  
25            # div(id = "ikelimoBox",  
26              box(width = 4, collapsible = TRUE,  
27                fileInput(  
28                  inputId = "failas_1",  
29                  label = "Å³kelkite duomenÅ³ failÅ³:",  
30                  multiple = TRUE,  
31                  accept = c(  
32                    "text/csv",  
33                    "text/comma-separated-values,text/plain",  
34                    ".csv",  
35                    ".xlsx"  
36                ),  
37                buttonLabel = "Å³kelti",  
38                placeholder = "NeÅ³keltas duomenÅ³ failas"),
```

server.R file example

```
1 # Define server logic for random distribution app ----
2 shinyserver(function(input, output) {
3
4   rvariables <- reactiveValues(duomenys = NULL, duomenys_first = NULL,
5                               isskirtys_final = NULL, select_edit_isskirtys = NULL,
6                               newduomenys_2 = NULL, final_nepateke_duomenys = NULL,
7                               nepateke_duomenys_2 = NULL, newduomenys = NULL,
8                               isskirtys = NULL, nepateke_duomenys_3 = NULL,
9                               summaryDF = NULL, nepateke_duomenys = NULL)
10
11   observeEvent(input$ikelimoBoxButton, {
12     # show_hide$a <- ifelse(input$ikelimoBoxButton %% 2 == 1,12,10)
13     toggle("ikelimoBox")
14   })
15   output$value <- renderText({
16     show_hide$a           # rv$value
17   })
18   observeEvent(input$button, {
19     toggle("hello")
20   })
21   failo_info <- reactive({input$failo_tipas1})
22   output$apie_faila <- renderUI({
23     if(!is.null(failo_info()) & failo_info() == ".csv"){
24       radioButtons(inputId = "sep",
25                   label = "skyrimo Åenklas:",
26                   choices = c(kablelis = ",",
27                               `kablataÅkis` = ";",
28                               tabuliatorius = "\t"),
29                   selected = ",")
30     }else{
31       numericInput(inputId = "sheetNr", label = "LenteleÅs nr.:",value = 1)
32     }
33   })
34   inputFile <- reactive({input$failas_1})
35   output$output_data <- renderUI({
36     if(!is.null(inputFile())){
37       if(input$skaityti_faila != 0){
```

Data file upload

Duomenų failo įkėlimas

Pagrindinės charakteristikos

Įskirčių radimas

Įkelkite duomenų failą:

Įkelti Duomenys.xlsx

Upload complete

Failo tipas:

XLSX

CSV

Lentelės nr.:

1

Nuskaityti failą

Įkelta duomenų lentelė

Search:

ID	Ekonominis veiklos kodas	Svoris	Pirm. apyvarta 2016 Q3	Darb. sk. 2016 Q3	PVM 2016 Q3	Pirm. apyvarta 2016 Q2	Darb. sk. 2016 Q2	PVM 2016 Q2	Pirm. apyvarta 2015 Q3	Darb. sk. 2015 Q3	PVM 2015 Q3
All	All	All	All	All	All	All	All	All	All	All	All
1	422200	1	3733020	345.51	2065401	2599084	370.64	2326159	2243000	6563587	525.62
2	412010	2.48979591836735	994696	22	1187516	499526	24.62	1040148			
3	431220	22.6753246755247	0	2		0	2	0			
4	432100	1	639601	75.64	693475	796926	76.51	803618	1192409	1170839	73.24

Main characteristics

Neteigiamų ir praleistų reikšmių šalinimas

Kintamieji, pagal kuriuos pašalinti netinkamas reikšmės (<0, =0 arba NA):

- ID
- Ekonominis veiklos kodas
- Svoris
- Pirm. apyvarta 2016 Q3
- Darb. sk. 2016 Q3
- PVM 2016 Q3
- Pirm. apyvarta 2016 Q2
- Darb. sk. 2016 Q2
- PVM 2016 Q2
- Pirm. apyvarta 2015 Q3
- Darb. sk. 2015 Q3
- PVM 2015 Q3

Pašalinti netinkamas reikšmės

[↶ Grįžti prie nevalytų duomenų](#)

Skyrimo ženklas .csv failo išsaugojimui:

- Kablelis
- Kabliataškis
- Tabulatorius

[⬇ Išsaugoti lentelę](#)

Main characteristics

	Pirm. apyvirta 2016 Q3	Darb. sk. 2016 Q3	PVM 2016 Q3	Pirm. apyvirta 2016 Q2	Darb. sk. 2016 Q2	PVM 2016 Q2	Pirm. apyvirta 2015 Q3	Darb. sk. 2015 Q3	PVM 2015 Q3
Min.	0	0	-343319	-330	0	-11830	0	0	2
1st Qu.	8443.75	7.28	50793	0	7.54	29399.75	211696	399152	43.42
Median	130427	22.89	237399	78236	23	172660	537503	794429	70.91
Mean	698391.9	48.83	851337.46	410366.03	48.39	723080.28	1113615.4	1924841.57	101.74
3rd Qu.	567357.5	58.7	698737.25	350173.75	57.38	545330	1117753	1788272	110.15
Max.	43939050	892.09	49025984	39190785	870.87	42895881	19674532	51082876	836.29

Duomenų pasiskirstymo tikrinimas

Kintančiojo ir histrogramos nustatymai

Pasirinkite kintamąjį:

Pirm. apyvirta 2016 Q3

Reikšmių rangu:



Skaiptelių kiekis:



Normalumo testas Šapiro-Vilko

H₀: Pasiskirstymo dėsnis yra normalusis

H₁: Pasiskirstymo dėsnis nėra normalusis

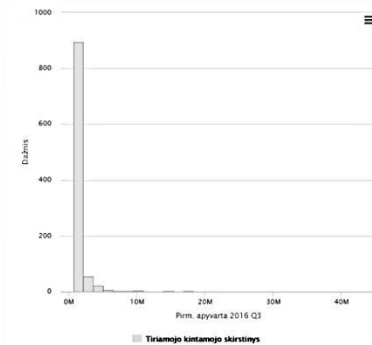
Reikšmingumo lygmuo α : 0.05

Statistika: 0.280

p-reikšmė: 0.000

Nulinė hipotezė yra atmesta, nes p-reikšmė yra žemesnė reikšmingumo lygmeniui. Todėl kintamasis nėra pasiskirstęs pagal normalųjį dėsnį.

Histograma Stabikampė diagrama



Main characteristics

Pagrindines kintamojo charakteristikos

	Min.	Mean	Max.	Stdev.	Size
Pirm. apyvara 2016 Q3	0	698391.9	43939050	2269211.69	996

Dviejų kintamųjų priklausomybė

Pasirinkite pirmąjį kintamąjį:

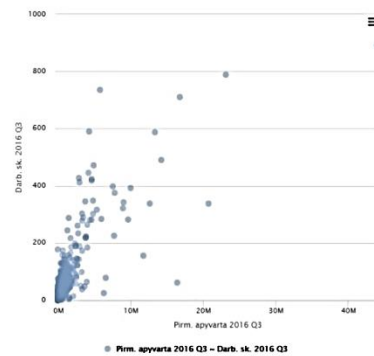
Pirm. apyvara 2016 Q3

Pasirinkite antrąjį kintamąjį:

Darb. sk. 2016 Q3

Koreliacijos koeficientas:	Pirsono	Spearmano
Įvertinta reikšmė:	0.778	0.858
H ₀ :	$\rho = 0$	$\rho = 0$
H _a :	$\rho \neq 0$	$\rho \neq 0$
Reikšmingumo lygmuo α :	0.05	0.05
Statistika:	39.028	23411863.976
p-reikšmė:	0.000	0.000

Pirsono ir Spearmano koreliacijos koeficiento nulios hipotezė yra atmetama, nes abiejų koeficientų p-reikšmės yra žemiaus reikšmingumo lygmenis.



Outliers detection

Analizės nustatymai

Kintamieji išskirtims rasti

Pasirinkite pagrindinį kintamąjį:

Pirm. apyvarta 2016 Q3

Pasirinkite papildomąjį kintamąjį (selective editing arba regresijos modeliui):

PVM 2016 Q3

Pasirinkite papildomąjį kintamąjį Hidiroglou-Berthelot metodui:

Pirm. apyvarta 2016 Q2

Metodai išskirtims rasti

Pasirinkite išskirčių radimo metodus:

- Pasirinktinis redagavimas (selective editing)
- Stebinio įtakos indeksas
- Kuko matas
- Standartizuotosios paklaidos
- DFBETAS matas
- Hidiroglou-Berthelot

Įmonių skaičius ekonominės veiklos grupėje

Pasirinkite minimalų įmonių skaičių:



✓ Taikyti

Outliers detection

Pasirinktinio redagavimo nustatymai


Epsilon parametras:

Įtraukti įmonių svarbos svorius

Stebinio įtakos indekso nustatymai

Stebinio įtakos indekso taisyklės parametras:

- 4
- 5
- 6

 Skaičiuoti

Outliers detection

Standartizuotosios paklaidos nustatymai

Standartizuotosios paklaidos taisyklės parametras:

- 1
- 2
- 3
- 4
- 5

DFBETAS mato nustatymai

DFBETAS mato taisyklės parametras:

- 2
- $2/\sqrt{2}$ (kai turima didelė duomenų aibė)

Outliers detection

Išskirčių atrinkimo nustatymai

Galutiniai išskirties kriterijai

Procentinių skirtumų slenkstis:

Pasirinkite teisingą išskirties pažymėtį:

- Pasirinktinis redagavimas
- Hidrologus-Berthelot
- Procentinis skirtumas

Lentelės su išskirtimis išsaugojimas

Skirimo ženklas .csv failo išsaugojimui:

- Kablelis
- Kablytaikis
- Tabulatorius

[Išsaugoti lentelę](#)

Nepatekusių duomenų nustatymai

Parametras priimtimumo intervalo ilgiui:

Skirimo ženklas .csv failo išsaugojimui:

- Kablelis
- Kablytaikis
- Tabulatorius

[Išsaugoti nepatekusius duomenis](#)

Lentelė su gautomis išskirtimis

Search:

Firm_apyvarta_2016_Q2	Darb_uk_2016_Q2	PVM_2016_Q2	Firm_apyvarta_2015_Q3	Darb_uk_2015_Q3	PVM_2015_Q3	Procentinis skirtumas	Parametras t.sel	Išskirtis pagal pasirinkimą	C	Išskirtis pagal HB	Išskirtis pagal proc. skirt.	Išskirtis pagal pasirinktus kriterijus
All	All	All	All	All	All	All	All	All				All
727493	85.54	807617	549058	614566	93.13	12.71	0.003	0	4	0	0	0
9409	66.09	459702				68.1	0.004	1	22	1	1	1
55890	6.74	487708	105264	1114372	18.38	21.62	0.036	0	9	0	1	0
383262	45.49	389917	765813	872510	51.48	1.61	0.028	0	20	0	0	0
192338	24.37	177292				0	0.036	0	9	0	0	0