Use of R in outliers detection

Audrius Taraila
audrius.taraila@stat.gov.lt

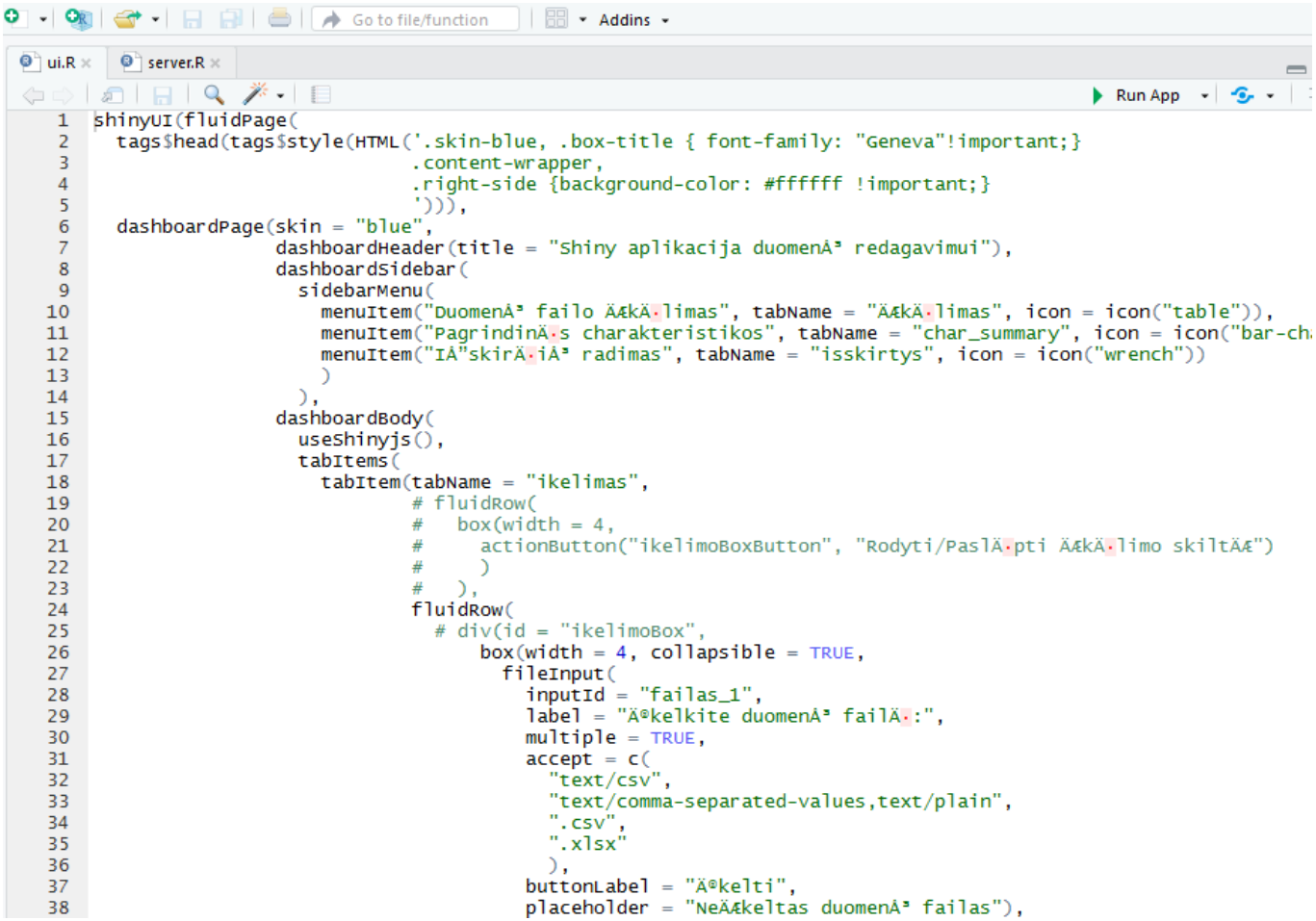
Statistics Lithuania

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

Shiny structure

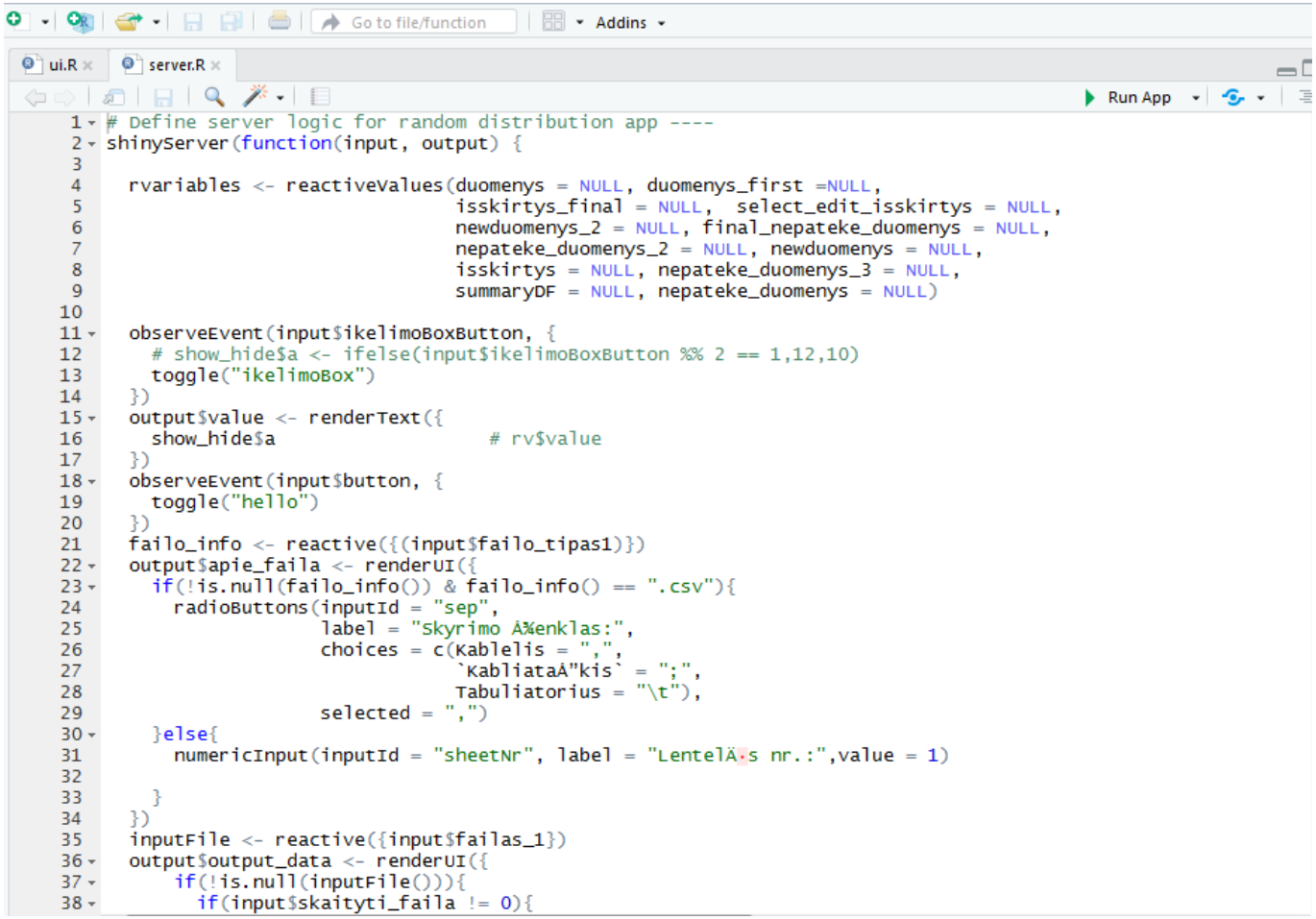
- Consists of two components
  - User interface object (ui.R file)
  - Server function (server.R file)

ui.R file example



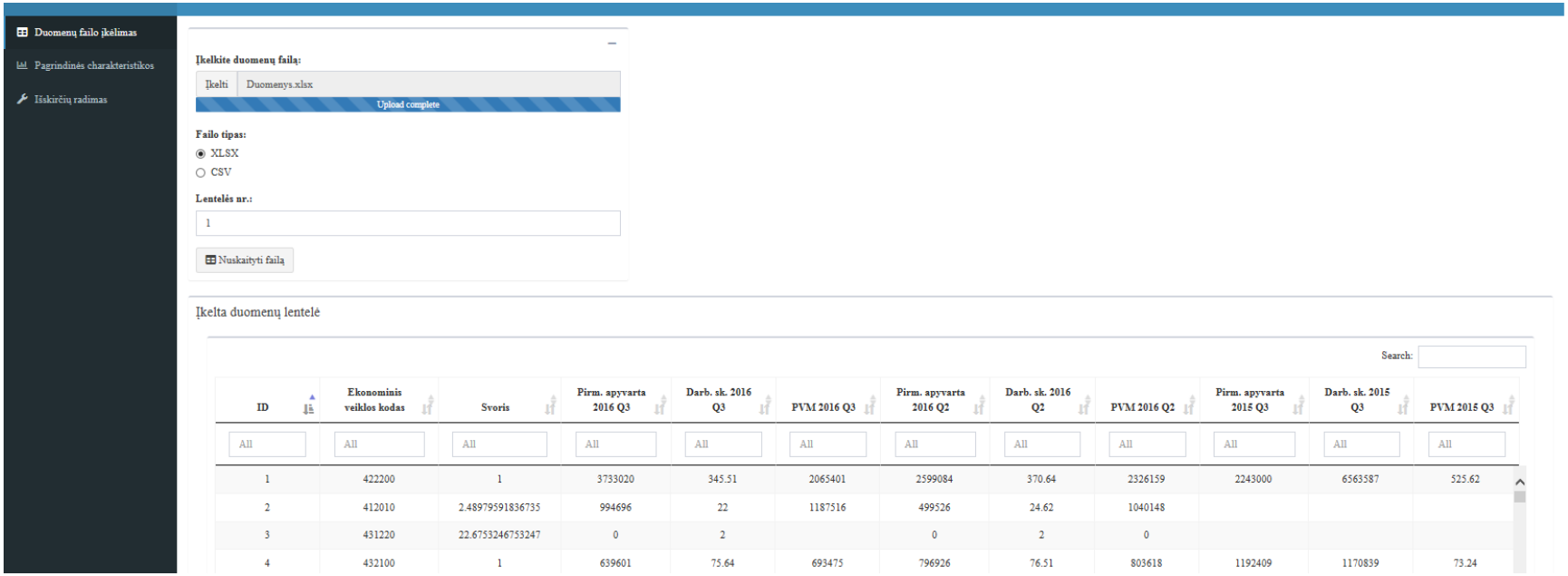

```
shinyUI(fluidPage(
  tags$head(tags$style(HTML('.skin-blue, .box-title { font-family: "Geneva"!important;}
                            .content-wrapper,
                            .right-side {background-color: #ffffff !important;}
                            '))),
  dashboardPage(skin = "blue",
                dashboardHeader(title = "Shiny aplikacija duomenÅ³ redagavimui"),
                dashboardSidebar(
                  sidebarMenu(
                    menuItem("DuomenÅ³ failo Ä¾kÄ·limas", tabName = "Ä¾kÄ·limas", icon = icon("table")),
                    menuItem("PagrindinÄ·s charakteristikos", tabName = "char_summary", icon = icon("bar-ch
                    menuItem("IÅ¡skirÄ·iÅ³ radimas", tabName = "isskirtys", icon = icon("wrench"))
                    )
                  ),
                dashboardBody(
                  useShinyjs(),
                  tabItems(
                    tabItem(tabName = "ikelimas",
                            # fluidRow(
                            #   box(width = 4,
                            #     actionButton("ikelimoBoxButton", "Rodyti/PaslÄ·pti Ä¾kÄ·limo skiltÄ¾")
                            #     )
                            #   ),
                            fluidRow(
                              # div(id = "ikelimoBox",
                                  box(width = 4, collapsible = TRUE,
                                    fileInput(
                                      inputId = "failas_1",
                                      label = "Ä®kelkite duomenÅ³ failÄ·:",
                                      multiple = TRUE,
                                      accept = c(
                                        "text/csv",
                                        "text/comma-separated-values,text/plain",
                                        ".csv",
                                        ".xlsx"
                                        ),
                                      buttonLabel = "Ä®kelti",
                                      placeholder = "NeÄ¾keltas duomenÅ³ failas"),
```

server.R file example

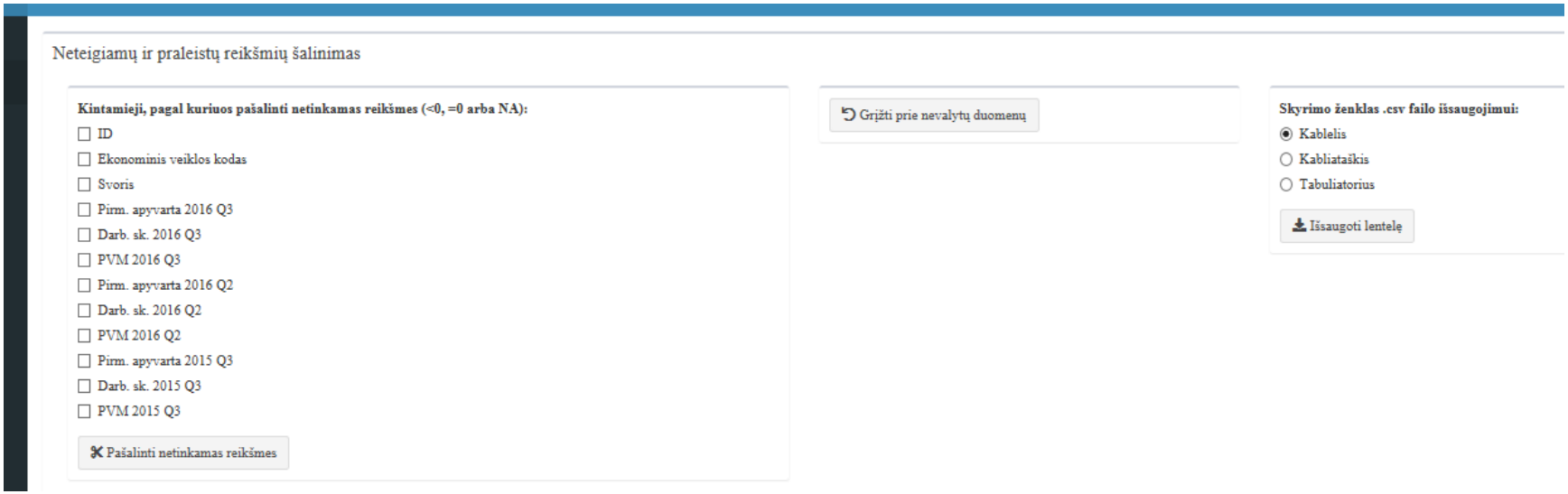



```r
# Define server logic for random distribution app ----
shinyServer(function(input, output) {

  rvariables <- reactiveValues(duomenys = NULL, duomenys_first =NULL,
                               isskirtys_final = NULL,  select_edit_isskirtys = NULL,
                               newduomenys_2 = NULL, final_nepateke_duomenys = NULL,
                               nepateke_duomenys_2 = NULL, newduomenys = NULL,
                               isskirtys = NULL, nepateke_duomenys_3 = NULL,
                               summaryDF = NULL, nepateke_duomenys = NULL)

  observeEvent(input$ikelimoBoxButton, {
    # show_hide$a <- ifelse(input$ikelimoBoxButton %% 2 == 1,12,10)
    toggle("ikelimoBox")
  })
  output$value <- renderText({
    show_hide$a                    # rv$value
  })
  observeEvent(input$button, {
    toggle("hello")
  })
  failo_info <- reactive({(input$failo_tipas1)})
  output$apie_faila <- renderUI({
    if(!is.null(failo_info()) & failo_info() == ".csv"){
      radioButtons(inputId = "sep",
                   label = "Skyrimo Ã¾enklas:",
                   choices = c(Kablelis = ",",
                               `KabliataÂ"kis` = ";",
                               Tabuliatorius = "\t"),
                   selected = ",")
    }else{
      numericInput(inputId = "sheetNr", label = "LentelÄ·s nr.:",value = 1)

    }
  })
  inputFile <- reactive({input$failas_1})
  output$output_data <- renderUI({
      if(!is.null(inputFile())){
        if(input$skaityti_faila != 0){
```

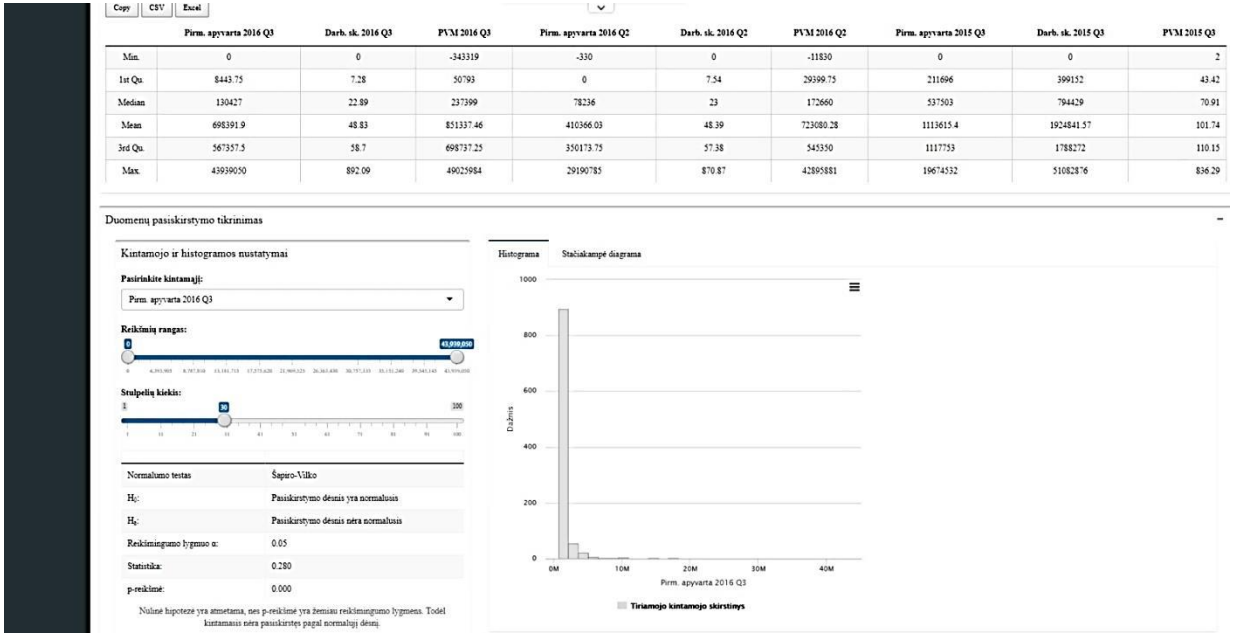
Data file upload



Main characteristics



Main characteristics



Copy CSV Excel

	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
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	545350	1117753	1788272	110.15
Max.	43939050	892.09	49025984	29190785	870.87	42895881	19674532	51082876	836.29

Duomenų pasiskirstymo tikrinimas

Kintamojo ir histogramos nustatymai

Pasirinkite kintamąjį:

Pirm. apyvarta 2016 Q3

Reikšmių rangas:

0 43,939,050

Stulpelių kiekis:

1 30 100

Normalumo testas	Šapiro-Vilko
H_0:	Pasiskirstymo dėsnis yra normalusis
H_1:	Pasiskirstymo dėsnis nėra normalusis
Reikšmingumo lygmuo α:	0.05
Statistika:	0.280
p-reikšmė:	0.000

Nulinė hipotezė yra atmetama, nes p-reikšmė yra žemiau reikšmingumo lygmens. Todėl kintamasis nėra pasiskirstęs pagal normalųjį dėsnį.

Histograma Stačiakampė diagrama

Main characteristics

Pagrindinės kintamojo charakteristikos

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

Dviejų kintamųjų priklausomybė

Pasirinkite pirmąjį kintamąjį:

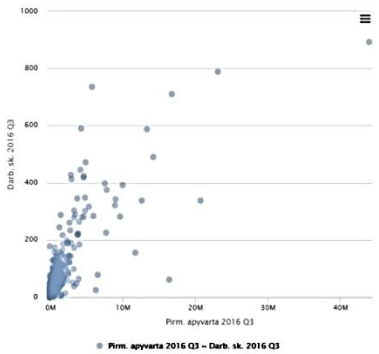
Pirm. apyvarta 2016 Q3

Pasirinkite antrąjį kintamąjį:

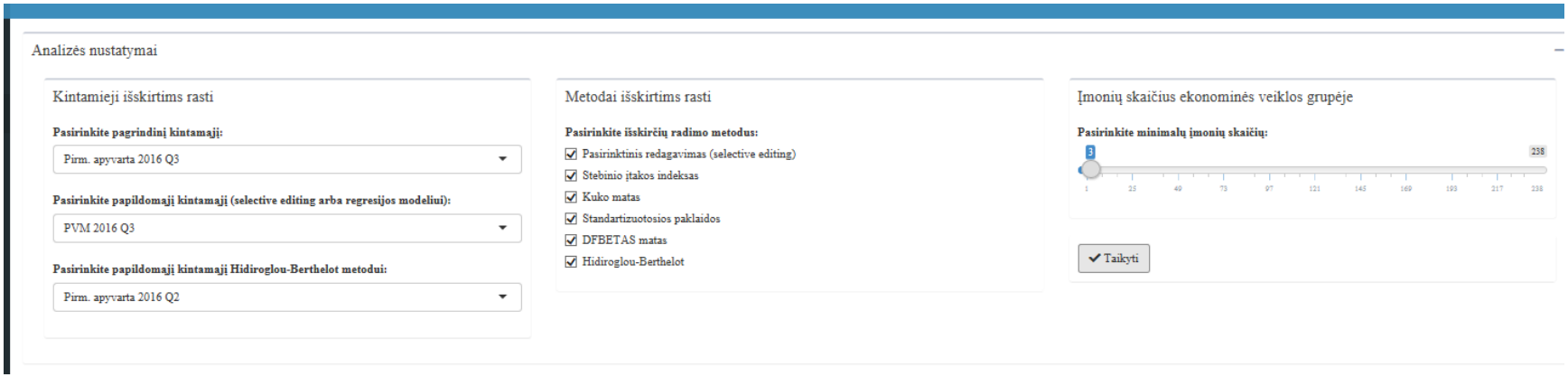
Darb. sk. 2016 Q3

Koreliacijos koeficientas:	Pirsono	Spirmano
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

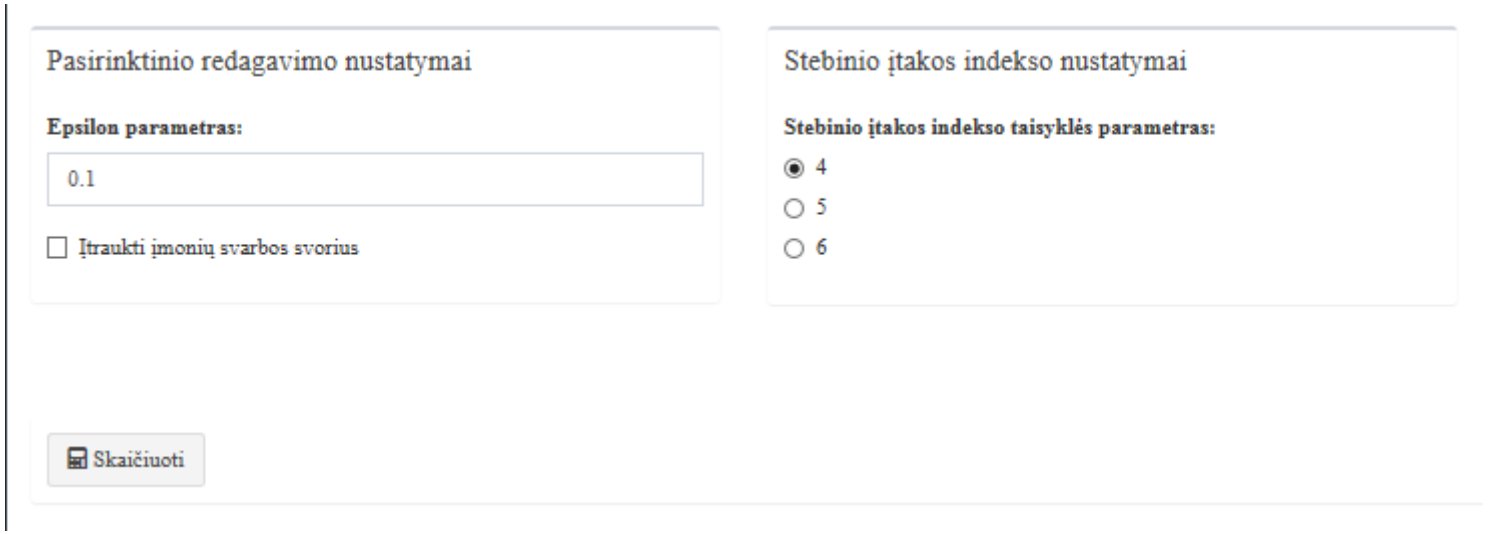
Pirsono ir Spirmano koreliacijos koeficiento nulinė hipotezė yra atmetama, nes abiejų koeficientų p-reikšmės yra žemiau reikšmingumo lygmens.



Outliers detection



Outliers detection



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

Pirm._apyvarta_2016_Q2	Darb._sk._2016_Q2	PVM_2016_Q2	Pirm._apyvarta_2015_Q3	Darb._sk._2015_Q3	PVM_2015_Q3	Procentinis skirtumas	Patametras t. sel	Išskirtis pagal pasir. redagavimą	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