УДК 004

ПРИМЕНЕНИЕ ЯЗЫКА ПРОГРАММИРОВАНИЯ R ДЛЯ СТАТИСТИЧЕСКОЙ ОБРАБОТКИ ДАННЫХ И РАБОТЫ С ГРАФИКОЙ В РЕШЕНИИ ПРИКЛАДНЫХ ЗАДАЧ

Величко Д.Д.

Магистр информационных систем и технологий,

Калужский государственный университет им. К.Э. Циолковского,

Калуга, Россия

Ткаченко А.Л.

к.т.н., доцент,

Калужский государственный университет им. К.Э. Циолковского,

Калуга, Россия

Кузнецова В.И.

к.п.н., доцент,

Калужский филиал Финансового университета при Правительстве Российской Федерации,

Калуга, Россия

Аннотация

Данная работа проводилась с целью рассмотрения различных методов визуализации числовых и текстовых данных с использованием языка **R** и прикладных пакетов, на основе данных о результатах Единого государственного экзамена (ЕГЭ) в школах условного города. Основным программным продуктом был язык **R** - язык программирования для статистической обработки данных и работы с графикой, а также свободная программная среда вычислений с открытым исходным кодом.

Для достижения поставленной цели применялись также дополнительные программные пакеты: igraph, TeachingDemos, tm.

Ключевые слова: визуализация данных, анализ данных, интеллектуальный анализ, таблицы, базы данных, гистограммы, диаграммы рассеяния, матрица классификации, диаграммы Кливленда, лепестковые диаграммы.

APPLICATION OF THE R PROGRAMMING LANGUAGE FOR STATISTICAL DATA PROCESSING AND WORKING WITH GRAPHICS IN SOLVING APPLIED PROBLEMS

Velichko D.D.
Master of Information Systems and Technology,
Kaluga State University named after K. E. Tsiolkovsky,
Kaluga, Russia

Tkachenko A.L.
candidate of Technical Sciences,
Kaluga State University named after K. E. Tsiolkovsky,
Kaluga, Russia

Kuznetsova V. I.
candidate of pedagogical Sciences,
Kaluga Branch of the Financial University under the Government of the Russian Federation,
Kaluga, Russia

Abstract

This work was carried out in order to consider various methods of visualizing numerical and textual data using the R language and application packages, based on data on the results of the Unified State Exam (USE) in schools of a conditional city.

The main software product was the R language, a programming language for statistical data processing and working with graphics, as well as a free open source computing software environment. To achieve this goal, additional software packages were also used: igraph, TeachingDemos, tm.
Keywords: data visualization, data analysis, data mining, tables, databases, histograms, scatterplots, classification matrix, cleveland diagrams, petal diagrams.

Данные, которые использовались для анализа Единого государственного экзамена в школах города, представлены на рисунке 1, для анализа использовался .csv формат данных. Данные представляют собой статистику средних балов ЕГЭ по 10 предметам и 43 школам [1].

Визуализация будет представлена в виде: диаграммы сравнения, диаграммы Кливленда и диаграммы размахов [2, 3]. Диаграмма сравнения позволит сделать некоторые предположения о взаимосвязи данных в диапазоне принимаемых значений. Диаграмма Кливленда аналогична по построению линейчатым диаграммам, но не включает в себя лишний параметр – ширину столбца. Диаграмма размахов, или «усатый ящик» (англ. box-whisker-plots), называется таки из-за его вида: точку или линию, соответствующую среднему положению совокупности данных, ограничивает прямоугольник («ящик»), длина которого соответствует одному из показателей разброса или точности оценки параметра [4].

Загрузим соответствующий файл .csv в таблицу и посмотрим данные (рисунок 2):

- ege=read.csv("school_ege.csv")
- plot(ege[2:10])

На полученных диаграммах представлены результаты попарного сравнения результатов экзаменов по предметам (координата x – результат по предмету, задаваемому в колонке, координата y – результаты по предмету, задаваемому в строке, точка представляет отдельную школу). По диаграммам

можно сделать некоторые предположения о взаимосвязи данных и диапазоне принимаемых значений.

Школа	Русский язык	Математика	Физика	Химия	Информатика	Биология	История	Английский язык	Обществознание	Литература
Школа 1	66.6	47.75	56.75	41	56.5	40	48.2	62.5	60.9	20
Школа 2	69.71	52.66	47	63	20	68.6	49	68.33	59.05	20
Школа 3	60.68	45.4	54.5	51.33	20	57.33	48.14	20	55.83	16
Школа 4	74.95	42	66.33	60.2	59	64.75	48.75	63	57.82	58
Школа 5	82.29	47	57.16	50	20	63	68.4	79.1	66.3	67.66
Школа 6	69.73	54.56	51.37	57.5	79	50.45	38	56.5	46.75	62
Школа 7	74.72	49.94	57.65	54.33	20	53.15	53.88	77.55	55.54	51
Школа 8	64.87	49.58	47.14	44	88	55	45	63.5	52	40
Школа 9	79.7	57.59	62.77	64.25	55.55	68.85	67.92	74	73.27	61.57
Школа 10	70.02	37.44	50.33	43.75	49.5	47.2	47.57	62.71	57.25	37
Школа 11	69.86	48.57	48.2	63.21	72	61.12	48.33	83	54.69	20
Школа 12	73.07	40.93	52.94	51.2	64	59.5	55.62	72.22	53.15	48.5
Школа 13	75.39	52.41	64.38	46.75	42	48	54.73	77.77	62.52	72
Школа 14	72.78	51.94	53.77	53.33	63	59.8	48.33	64.88	57.88	64
Школа 15	72.04	45.35	47.17	48.28	55	53.86	61.46	60.8	60.73	41.8
Школа 16	70.21	49.08	45.5	55.5	61	61.2	53.57	51	59.07	66.4
Школа 17	73.66	47.04	50.25	35.75	39.8	39.55	53.12	74.1	55.61	49
Школа 18	81.55	58.12	69	71.83	68	71	64.88	62.5	63.86	61.66
Школа 19	69.2	36.86	48.31	55.66	20	41.62	47.22	57.4	50.36	62.33
Школа 20	66.51	46.92	49.85	49.5	58	49.62	40	61.33	50.8	20
Школа 21	72.56	59.2	49.83	52.75	60.5	60.5	52.62	88	56.94	39.5
Школа 22	87.06	56.75	70	56.69	74	64.69	66.14	84.6	68.06	73
Школа 23	76.04	48.27	55.83	50	48	42.83	54.33	91	56.2	20
Школа 24	75.07	39.9	56.33	61.9	20	65.77	60.43	68.5	60	20
Школа 25	65	41.83	58	43	70	49	53	66	49.64	40.5
Школа 26	71.06	41.27	42.66	61	20	39.33	35.33	73	48.85	34.5
Школа 27	73.23	58.82	56.16	61.5	58.25	67	52.66	64	60.72	53
Школа 28	61.43	48.86	55.66	8	20	43.33	37.5	20	49.6	20
Школа 29	80.13	54.12	57.25	61.62	86.25	69.57	52.71	58.2	58.21	57
Школа 30	70.71	58.25	54	20	68	42.33	20	67	52.71	20
Школа 31	76	29	36	44	20	51.66	20	20	51.25	20
Школа 32	46.5	25	20	20	20	34	25	20	46	20
Школа 33	66.63	41.12	42.08	62.5	57	56.33	43	53	49.44	43
Школа 34	76.45	54.86	62.54	72	65.71	69.54	50.36	75.9	54.35	55.25
Школа 35	80.18	60.18	59.7	63	56.25	62.1	57.8	83	65.55	78
Школа 36	57.91	33.4	63.5	20	20	39.5	49.66	20	56.28	20
Школа 37	78.33	62	64.62	66	20	57	63.18	75.73	65.95	54.33
Школа 38	72.16	47.96	53.88	60.5	65	49.25	54	74.2	58.23	20
Школа 39	64.56	34.42	47.36	59.33	56	59.75	46.4	59	53.84	50.5
Школа 40	69.27	44.54	51.77	48.33	64.25	53.5	50	50.33	55.21	20
Школа 41	72.07	50.55	52.5	66.8	51.8	56.62	41	20	57.08	20
Школа 42	65.88	25	36	55	7	55.5	20	20	46.5	47
Школа 43	67.42	46	46.66	55.33	50	59	49.5	82	62	20

Рисунок 1 – Данные о среднем балле ЕГЭ по предметам в школах города. возможности. Подготовлено коллективом авторов.

Так, из диаграммы можно предположить, что результаты экзаменов между биологией-химией могут быть описаны статистически значимой линейной регрессией.

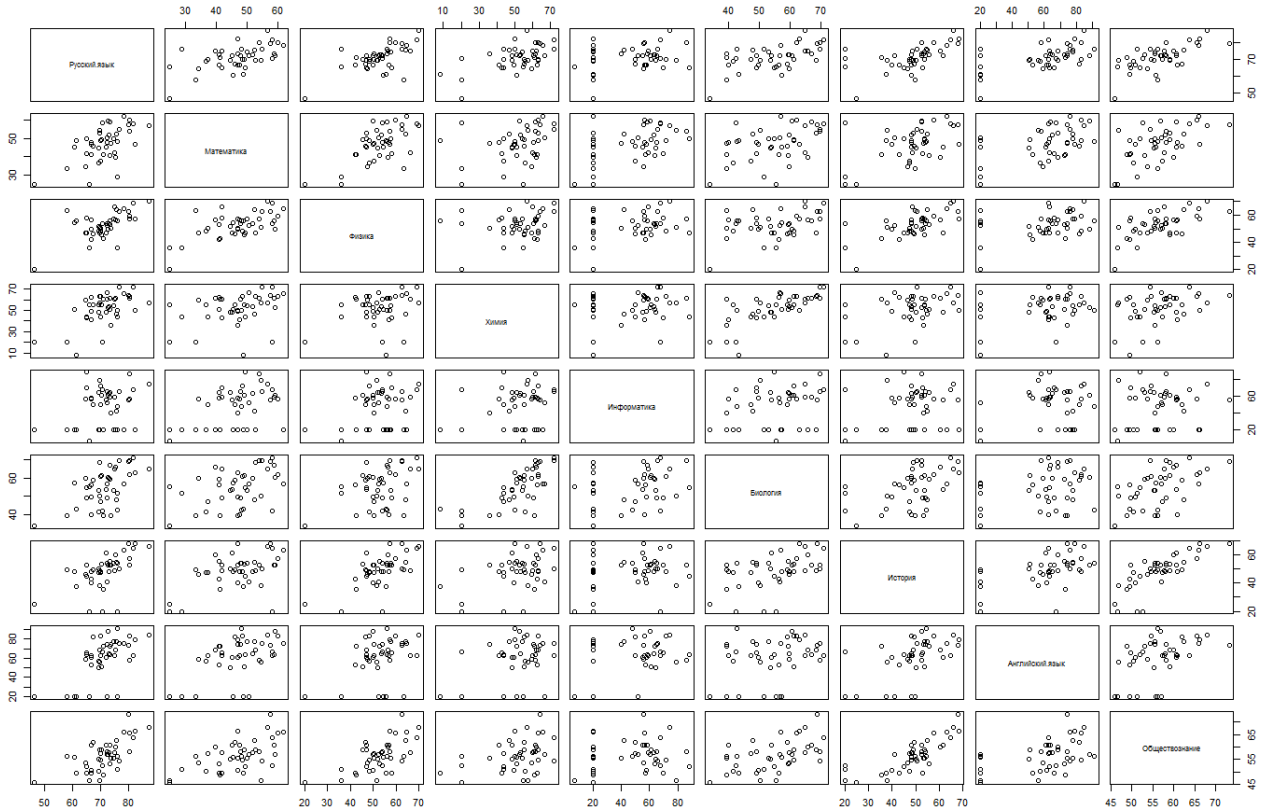

Рисунок 2 - Диаграмма сравнения. Подготовлено коллективом авторов.

Посмотрим результаты ЕГЭ для отдельно взятой школы:

- N=2 (номер образовательного учреждения)
- s=ege[N,2:10]
- s=t(s)
- dotchart(s,xlab = "Средний балл ЕГЭ по предмету")

Из такой диаграммы (рисунок 3) можно сделать ряд полезных выводов, например, то что лучше всего во 2 школе сдают английский язык, биологию, русский язык, в среднем на 70 баллов.

Также из диаграммы следует вывод, что средний балл в школе по большей части предметов превышает 45 баллов (рисунок 4).

Большая часть ребят получают баллы в диапазоне от 50 до 70 баллов.

Проведем кластеризацию (4 кластера):

- distances = dist(ege[2:10], method = "euclidean")
- hcluster = hclust(distances, method = "ward.D")
- clusterGroups = cutree(hcluster, k = 4)
- ege$clust =clusterGroups

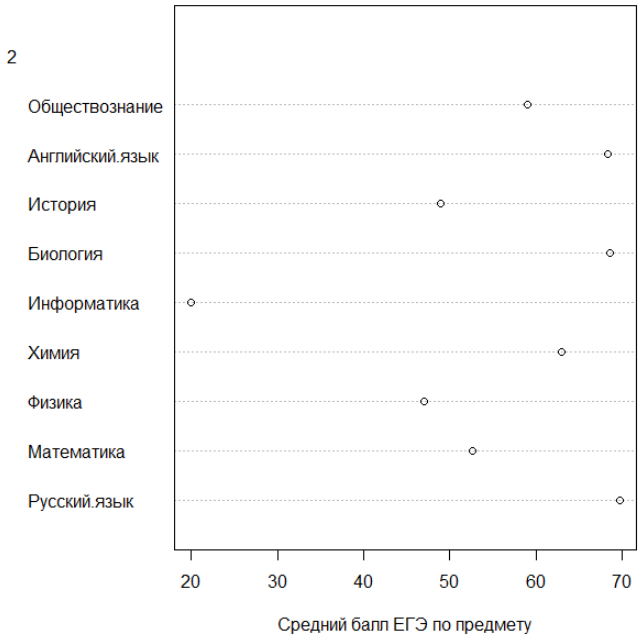

Рисунок 3 - Диаграмма Кливленда. Подготовлено коллективом авторов.

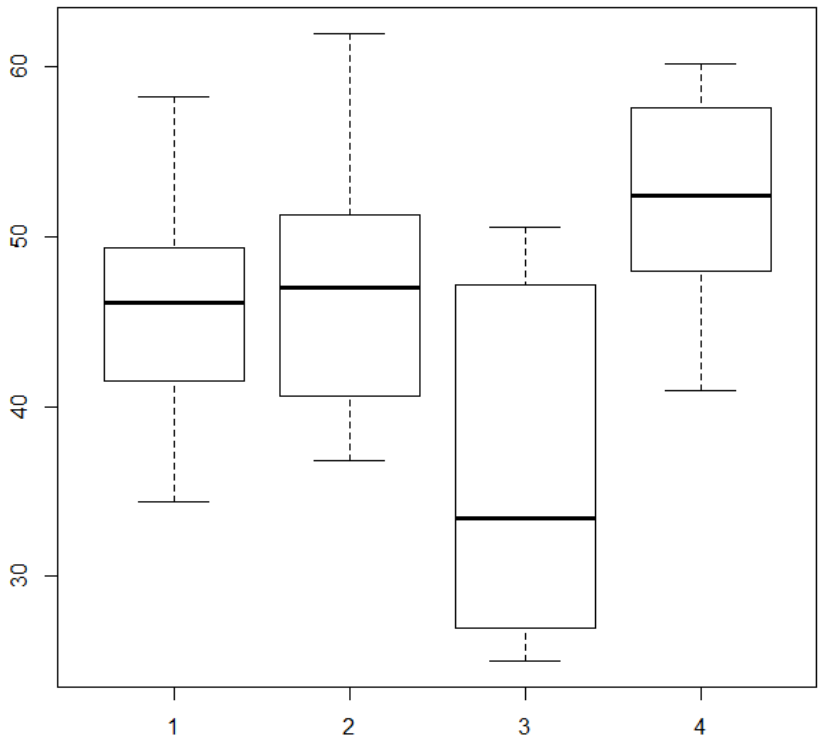

Рисунок 4 - Диаграмма размахов. Подготовлено коллективом авторов.

Теперь сравним результаты ЕГЭ по математике для разных кластеров:

- boxplot(ege$Математика ~ ege$clust)

Толстая черная линия – это медиана. «Усы» тянутся к самому максимальному и минимальному значениям (рисунок 4).

Наблюдения, находящиеся за границами «усов», могут быть выбросами. Несмотря на это, следует внимательно относится к такого рода нестандартным явлениям, так как они могут быть нормальными.

Таким образом мы рассмотрели несколько инструментов языка R для статистической обработки данных и работы с графикой на примере средних баллов ЕГЭ по школам. Из «сухого» набора данных можно сделать ряд полезных выводов, что поможет принять качественные управленческие решения, основанные на данных.

Библиографический список:

1. Имитационное моделирование демографических показателей роста и убыли наслеления / А. Л. Ткаченко, О. М. Лыкова, Е. И. Шаронов, В. И. Кузнецова // Modern Economy Success. – 2021. – № 3. – С. 110-116.
2. Кондрашова Н.Г. Информация и ее применение в ходе управления проектами // Дневник науки. 2020. № 12 (48). С. 50.
3. Суслякова, О. Н. Использование системы Deductor для интеллектуального анализа развития страхового рынка и построения прогноза / О. Н. Суслякова, А. Л. Ткаченко, С. В. Пономарев // Финансовая экономика. – 2019. – № 4. – С. 94-98.
4. Ткаченко, А. Л. Применение искусственного интеллекта в управленческих информационных системах / А. Л. Ткаченко // Развитие управленческих и информационных технологий, их роль в региональной экономике : материалы II Международной открытой научно-практической конференции, Калуга, 21–22 апреля 2016 года / Под

редакцией: Пироговой Т.Э., Швецовой С.Т., Орловцевой О.М. – Калуга: ООО "ТРП", 2016. – С. 147-153.

Оригинальность 75%