«به نام خدا»

تکلیف دوم – سوال پنجم – مرضیه علیدادی – 9631983

(کد های مربوط، در دو فرمت py. و ipynb. ضمیمه شده اند. – دیتاست اصلاح شده Diabetes_cleared.csv نیز ضمیمه شده است.)

.5

a) تحلیل ها در سوال 3 صورت گرفته است.

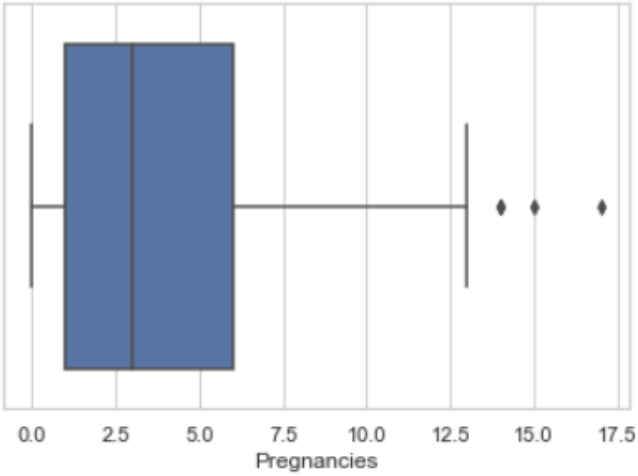
b) با استفاده از قابلیت های موجود و با نوشتن تابعی به کمک IQR، داده های پرت هر ستون را شناسایی کردم. همچنین، برای هر ستون نمودار جعبه‌ای رسم کردم.

برای مثال، برای ستون Pregnancies داریم:

- داده های پرت:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
88	15.0	136.0	70.0	32.0	110.000000	37.1	0.153	43.0	1.0
159	17.0	163.0	72.0	41.0	114.000000	40.9	0.817	47.0	1.0
298	14.0	100.0	78.0	25.0	184.000000	36.6	0.412	46.0	1.0
455	14.0	175.0	62.0	30.0	163.513514	33.6	0.212	38.0	1.0

- نمودار جعبه ای:



(برای بقیه ی متغیر ها، در کد ضمیمه شده می توان این اطلاعات را مشاهده کرد.)

پس از حذف داده های پرت تشخیص داده شده، تعداد سطر های دیتاست، برابر 338 تا شد. دیتاست اولیه، 502 تا سطر داشت. بنابراین، یعنی تعداد سطر های حذف شده با این روش، برابر 164 تا است.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6.0	148.0	72.0	35.000000	163.513514	33.6	0.627	50.0	1.0
1	1.0	85.0	66.0	29.000000	163.513514	26.6	0.351	31.0	0.0
2	8.0	183.0	64.0	28.904494	163.513514	23.3	0.672	32.0	1.0
3	1.0	89.0	66.0	23.000000	94.000000	28.1	0.167	21.0	0.0
5	5.0	116.0	74.0	28.904494	163.513514	25.6	0.201	30.0	0.0
...	...	...	...	...	...	...	...	...	...
497	2.0	81.0	72.0	15.000000	76.000000	30.1	0.547	25.0	0.0
498	7.0	195.0	70.0	33.000000	145.000000	25.1	0.163	55.0	1.0
499	6.0	154.0	74.0	32.000000	193.000000	29.3	0.839	39.0	0.0
500	2.0	117.0	90.0	19.000000	71.000000	25.2	0.313	21.0	0.0
501	3.0	84.0	72.0	32.000000	163.513514	37.2	0.267	28.0	0.0

338 rows × 9 columns

c) با استفاده از روش zScore داده های پرت را حذف کردم. تعداد سطر های دیتاست، برابر 460 تا شد. دیتاست اولیه، 502 تا سطر داشت. بنابراین، یعنی تعداد سطر های حذف شده با این روش، برابر 42 تا است.

در قسمت قبل هم مشاهده شد که با استفاده از روش IQR، تعداد سطر های حذف شده، برابر 164 تا بود.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6.0	148.0	72.0	35.000000	163.513514	33.6	0.627	50.0	1.0
1	1.0	85.0	66.0	29.000000	163.513514	26.6	0.351	31.0	0.0
2	8.0	183.0	64.0	28.904494	163.513514	23.3	0.672	32.0	1.0
3	1.0	89.0	66.0	23.000000	94.000000	28.1	0.167	21.0	0.0
5	5.0	116.0	74.0	28.904494	163.513514	25.6	0.201	30.0	0.0
...	...	...	...	...	...	...	...	...	...
497	2.0	81.0	72.0	15.000000	76.000000	30.1	0.547	25.0	0.0
498	7.0	195.0	70.0	33.000000	145.000000	25.1	0.163	55.0	1.0
499	6.0	154.0	74.0	32.000000	193.000000	29.3	0.839	39.0	0.0
500	2.0	117.0	90.0	19.000000	71.000000	25.2	0.313	21.0	0.0
501	3.0	84.0	72.0	32.000000	163.513514	37.2	0.267	28.0	0.0

460 rows × 9 columns