به نام خدا

تمرین چهارم درس مبانی داده‌کاوی

امیررضا حسینی ۹۸۲۰۳۶۳

سؤالات تئوری

سؤال ۱) در این سؤال تعداد سبدها را به‌صورت پیش‌فرض برابر با ۳ قرار می‌دهیم:

Equal-depth (frequency) partitioning:

Bin1: [5, 10, 11, 13]
Bin2: [15, 35, 50, 55]
Bin3: [72, 92, 204, 215]

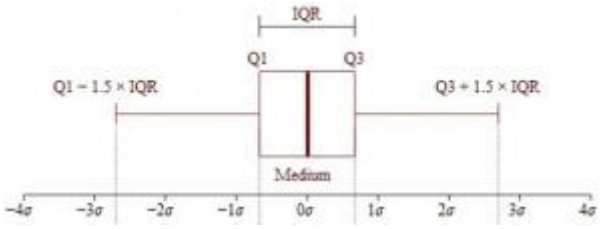
Equal-width (distance) partitioning:

w = (max – min) / (no of bins) = (215-5)/3=70

[min + w], [min + 2w] …. [min + nw] => [min + w], [min + 2w],[min+3w]

Bin1: [5, 10, 11, 13, 15, 35, 50, 55, 72]
Bin2: [92]
Bin3: [204, 215]

سؤال ۲)



محاسبه مقادیر چارک‌ها:

$$Q1 = \frac{8+8}{2} = 8$$

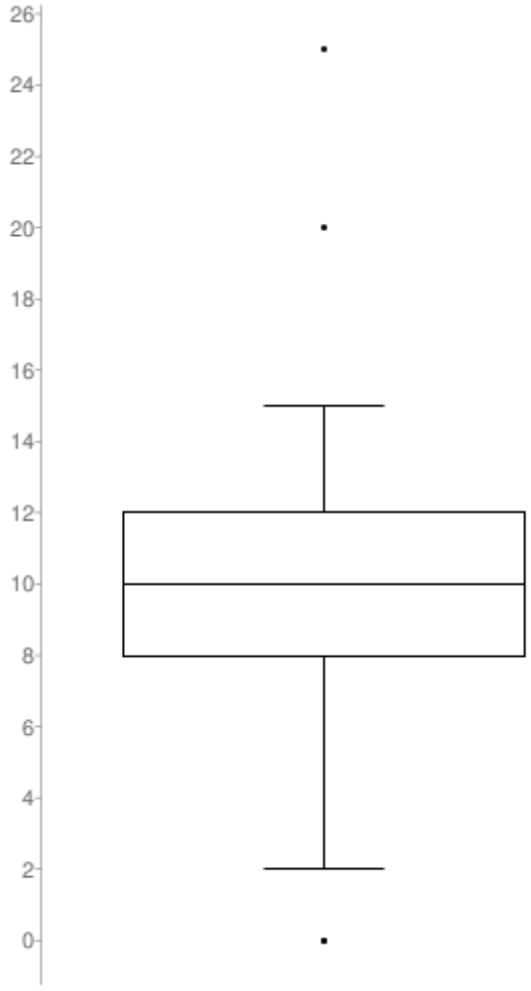
Median= 10

$$Q3 = \frac{12+12}{2} = 12$$

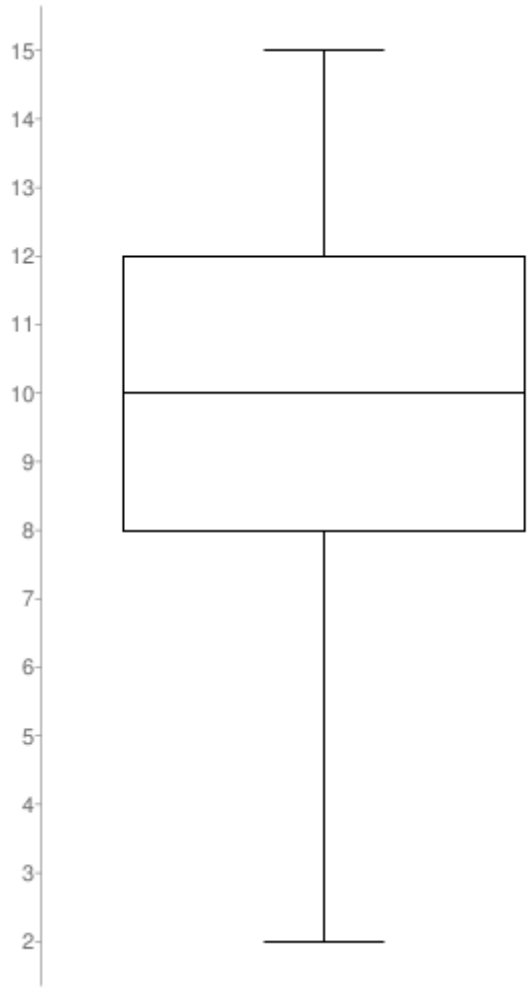
$$IQR = 12 - 8 = 4$$

(Q1 - 1.5×IQR = 2 | Q3 + 1.5×IQR = 18) => outliers = { **0, 0, 20, 25 }**

نمودار **box plot** قبل از حذف کردن داده‌های پرت:



نمودار **box plot** بعد از حذف کردن داده‌های پرت:



سؤال ۳)

ابتدا داده‌ها را سورت می‌کنیم که در اینجا در ابتدای کار سورت شده هستند. سپس داده‌ها را به binهایی با عمق یکسان (۳) تقسیم می‌کنیم:

Bin1: 13, 15, 16	Bin2: 16, 19, 20	Bin3: 20, 21, 22
Bin4: 22, 25, 25	Bin5: 25, 25, 30	Bin6: 33, 33, 35
Bin7: 35, 35, 35	Bin8: 36, 40, 45	Bin9: 46, 52, 70

سپس برای هموارسازی داده‌ها از دو روش زیر استفاده می‌کنیم:

Smoothing by **bin means**:

Bin1: 14.67, 14.67, 14.67
Bin2: 18.33, 18.33, 18.33
Bin3: 21, 21, 21
Bin4: 24, 24, 24
Bin5: 26.67, 26.67, 26.67
Bin6: 33.67, 33.67, 33.67
Bin7: 35, 35, 35
Bin8: 40.33, 40.33, 40.33
Bin9: 56, 56, 56

Smoothing by **bin boundaries**:

Bin1: 13, 16, 16	Bin2: 16, 20, 20	Bin3: 20, 22, 22
Bin4: 22, 25, 25	Bin5: 25, 25, 30	Bin6: 33, 33, 35
Bin7: 35, 35, 35	Bin8: 36, 36, 45	Bin9: 46, 46, 70

روش‌های پیداکردن داده‌های پرت: علاوه بر روشی که در سؤال دوم استفاده شد، یکی دیگر از روش‌های مؤثر برای حذف این نوع داده‌ها روش z-score و clustering است. در روش z-score ابتدا داده‌ها را وارد فضای جدیدی می‌بریم که با استفاده از میانگین و انحراف معیار داده‌هایی را که مقدار قدرمطلق آنها از ۳ بیشتر است را حذف می‌کنیم و در روش خوشه‌بندی نیز داده‌هایی که پس از خوشه‌بندی از خوشه‌ها فواصل زیادی دارند را داده پرت تشخیص می‌دهیم و سپس به حذف‌کردن آنها اقدام می‌کنیم. این دو روش unsupervised هستند.
روش دیگری که می‌توان برای تحقق این هدف ارائه داد وجود روش‌هایی با استفاده از دخالت انسان می‌باشد که بر حسب دانایی انسان از کسب‌وکار، مشخص کند که آیا این مقادیر برای داده‌ها پرت حساب می‌شوند یا نه.

روش‌های دیگری که برای smooth کردن این داده‌ها وجود دارد شامل تمامی روش‌هایی است که از آنها در سبد بندی استفاده می‌شود. مثلاً جایگذاری میانه، مد، boundary و غیره به جای داده‌های هر سبد.
روش‌های دیگری مانند رگرسیون و حذف نویز سیگنالی نیز می‌توان استفاده کرد که با یک lowpass filter انجام می‌شود.
همچنین از روش‌های دسته‌بندی سلسله‌مراتبی نیز می‌توان برای تحقق این امر استفاده کرد.