به نام خدا

مرضیه علیدادی – ۸۱۰۱۰۱۲۳۶

تکلیف ۱ داده کاوی

سوال ۱:

- مجموعه داده‌ی دما: داده‌ی Numeric از نوع پیوسته است. اگر با واحد سانتی‌گراد یا فارنهایت بیان شود، Interval-scaled است. و اگر با واحد کلوین بیان شود، Ratio-scaled است.
- مجموعه داده‌ی رطوبت: داده‌ی Numeric از نوع پیوسته و Quantity است.

- مجموعه داده‌ی جنسیت (مرد یا زن): داده‌ی Nominal از نوع باینری متقارن (Symmetric) است.
- مجموعه داده‌ی قد (سانتی‌متر): داده‌ی Numeric از نوع پیوسته و Ratio-scaled است.

- مجموعه داده‌ی سن (برحسب سال): داده‌ی Numeric از نوع پیوسته و Quantity است.
- مجموعه داده‌ی مربوط به اینکه آیا فرد دارای شرایط پزشکی است یا خیر: داده‌ی Nominal از نوع باینری است. بسته به اینکه آیا جواب بله و خیر، به میزان یکسان اهمیت دارند یا خیر، به ترتیب، از نوع متقارن (Symmetric) یا نامتقارن (Asymmetric) است. احتمالاً اینجا از نوع نامتقارن خواهدبود.

- مجموعه داده‌ی نوع خودرو (سدان، شاسی بلند یا کامیون): داده‌ی Nominal است.
- مجموعه داده‌ی اسب بخار: داده‌ی Numeric از نوع پیوسته و Interval-scaled است.
- مجموعه داده‌ی مصرف سوخت در صد کیلومتر: داده‌ی Numeric از نوع پیوسته و Ratio-scaled است.

- مناسب‌ترین نمودار برای مصور سازی رابطه‌ی بین مجموعه داده‌های دما و رطوبت، نمودار scatter plot است. این نمودار برای تشخیص رابطه‌های غیرخطی مناسب است. این نمودار برای تشخیص رابطه‌ی بین دو متغیر پیوسته، کارآمدترین است.
- مناسب‌ترین نمودار برای مصور سازی رابطه‌ی بین مجموعه داده‌های جنسیت و قد، نمودار histogram است. به این شکل که محور قد را به تعدادی bin تقسیم کنیم، و طول هر ستون رسم شده در هر bin نشان‌گر فراوانی افراد دارای آن قد باشد. و همچنین، هر ستون با دو رنگ متناظر با جنسیت افراد، رنگ شده باشد. تا فراوانی دو جنسیت در هر bin مشخص باشد و همچنین پراکندگی هر جنسیت در binها قابل برداشت باشد. محدودیتی که ایجاد می‌شود، این است که سن دقیق افراد را نمی‌توانیم ببینیم؛ زیرا داده‌ها به چند رده‌ی سنی (bin) تقسیم شده‌اند.
- مناسب‌ترین نمودار برای مصور سازی رابطه‌ی بین مجموعه داده‌های سن و اینکه آیا فرد دارای شرایط پزشکی است یا خیر، نمودار boxplot است. با توجه به این‌که دارای شرایط پزشکی بودن دارای دو مقدار بله یا خیر است، دو نمودار boxplot خواهیم داشت. یکی متناظر با افرادی که دارای شرایط پزشکی هستند، و دیگری متناظر با افرادی که دارای شرایط پزشکی نیستند. و محور دیگر مربوط به سن خواهد بود. در این دو نمودار مشخص خواهدشد که در هر یک از این دو دسته، عموماً افراد دارای چه سنی هستند. میانگین سنی و بازه‌ی سنی افراد مشخص خواهد شد. به کمک این نمودار، داده‌های با تعداد بالا قابل مصورسازی هستند. اما با توجه به این‌که خلاصه سازی انجام می‌دهد، جزئیات قابل برداشت نخواهدبود.
- مناسب‌ترین نمودار برای مصور سازی رابطه‌ی بین مجموعه داده‌های نوع خودرو، اسب بخار و مصرف سوخت در صد کیلومتر، نمودار parallel coordinates است. هر یک از نوع خودروها را با یک رنگ نشان می‌دهیم. و دو axe موازی، یکی متناظر با اسب بخار و دیگری متناظر با مصرف سوخت خواهیم داشت. به این شکل بهترین مصورسازی و مقایسه بین انواع خودروها از نظر اسب بخار و مصرف سوخت در صد کیلومتر را خواهیم‌داشت. با توجه به اینکه داده‌ی مربوط به نوع خودرو از نوع nominal است و دو داده‌ی دیگر، دارای مقادیر پیوسته هستند،

این نمودار انتخاب مناسبی برای مصورسازی روابط به این ۳ مجموعه داده است. این نمودار به سادگی روابط به نشان می‌دهد. از جمله مشکلاتی که می‌تواند وجود داشته باشد، همپوشانی خطوط در مقادیر رایج است. همچنین، مقایسه‌ی بین خودروها به صورت کیفی و نه کمی قابل انجام است. رابطه‌ی بین مجموعه داده‌ها را به شکلی که بتوان گفت رابطه‌ای خطی یا غیر خطی است، نمی‌توان تشخیص داد.

سوال ۲:

مجموعه داده‌ها به ترتیب:

۵، ۶، ۷، ۸، ۹، ۱۰، ۱۱، ۱۲، ۱۳، ۱۴، ۱۵، ۱۶، ۱۷، ۱۸، ۱۹، ۲۰، ۲۱، ۲۲، ۲۳، ۲۴، ۲۵

- روش عرض مساوی (با پنج bin):

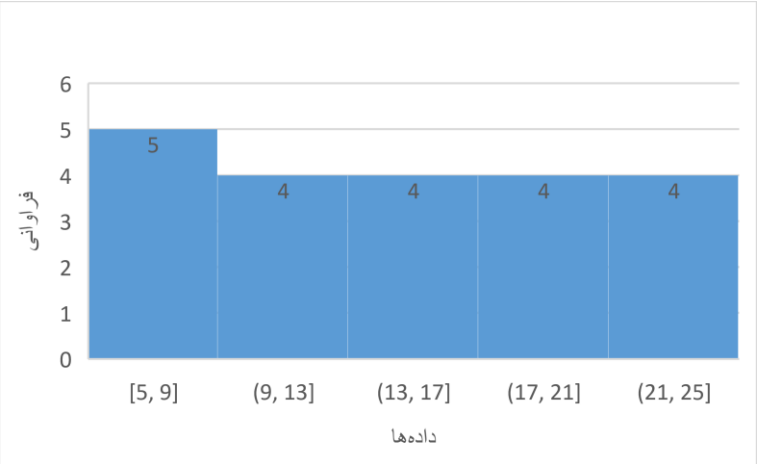
۱. (۵، ۶، ۷، ۸، ۹)

۲. (۱۰، ۱۱، ۱۲، ۱۳)

۳. (۱۴، ۱۵، ۱۶، ۱۷)

۴. (۱۸، ۱۹، ۲۰، ۲۱)

۵. (۲۲، ۲۳، ۲۴، ۲۵)

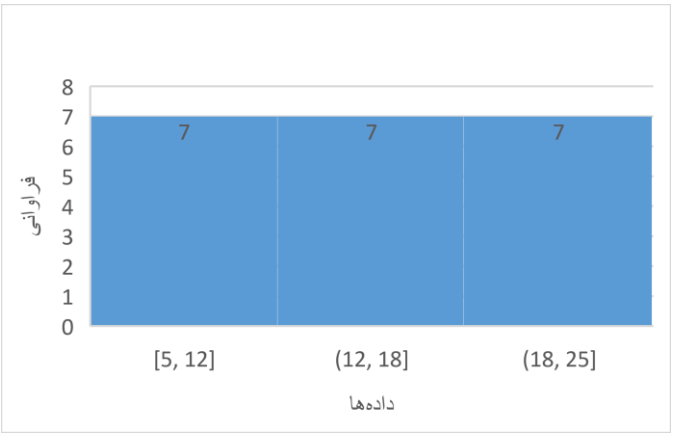


- روش عمق مساوی (با سه bin):

۱. (۵، ۶، ۷، ۸، ۹، ۱۰، ۱۱)

۲. (۱۲، ۱۳، ۱۴، ۱۵، ۱۶، ۱۷، ۱۸)

۳. (۱۹، ۲۰، ۲۱، ۲۲، ۲۳، ۲۴، ۲۵)



در مجموعه داده‌ی مربوط به این سوال، با توجه به اینکه داده‌ها با فاصله‌های یکسان قرار گرفته‌اند(تمام اعداد طبیعی از ۵ تا ۲۵، با فاصله‌های یکی یکی)، این دو روش دسته بندی تفاوت قابل توجهی ایجاد نمی‌کند. بسته به عمق یا عرضی که برای binning در نظر می‌گیریم، هر دو روش، داده‌ها را به تعدادی bin با عرض و عمق مساوی تقسیم می‌کند. (تقریباً مساوی – زیرا ممکن است تعداد داده‌ها، بر عدد انتخاب شده برای عمق یا عرض، بخشپذیر نباشد.)

سوال ۳:

۱.

- مجموعه داده‌ی سن به صورت مرتب:

۲۳، ۲۷، ۳۱، ۳۲، ۳۵، ۳۸، ۴۱، ۴۵، ۴۷، ۵۲

- **میانه:** 36.5

عددی که داده‌ها را به دو بخش با تعداد مساوی تقسیم می‌کند، همان میانه است. با توجه به این‌که تعداد داده‌ها زوج است، میانگین دو عدد وسط، به عنوان میانه در نظر گرفته می‌شود:

$$\frac{35 + 38}{2} = \frac{73}{2} = 36.5$$

- **چارک اول:** 31

میانه، داده‌ها را به دو بخش تقسیم می‌کند. میانه‌ی بخش اول، همان چارک اول کل داده‌هاست. داده‌های بخش اول:

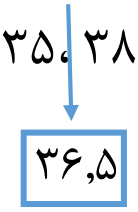
۲۳، ۲۷، ۳۱، ۳۲، ۳۵

- **چارک سوم:** 45

میانه، داده‌ها را به دو بخش تقسیم می‌کند. میانه‌ی بخش دوم، همان چارک سوم کل داده‌هاست. داده‌های بخش دوم:

۳۸، ۴۱، ۴۵، ۴۷، ۵۲

۲۳، ۲۷، ۳۱، ۳۲، ۳۵، ۳۸، ۴۱، ۴۵، ۴۷، ۵۲



- مجموعه داده‌ی درآمد به صورت مرتب:

۲۵۰۰۰، ۲۷۰۰۰، ۳۰۰۰۰، ۳۵۰۰۰، ۳۸۰۰۰، ۴۰۰۰۰، ۴۲۰۰۰، ۴۵۰۰۰، ۵۰۰۰۰، ۵۵۰۰۰

  - **میانه:** 39000

عددی که داده‌ها را به دو بخش با تعداد مساوی تقسیم می‌کند، همان میانه است. با توجه به این‌که تعداد داده‌ها زوج است، میانگین دو عدد وسط، به عنوان میانه در نظر گرفته می‌شود:

$$\frac{38000 + 40000}{2} = \frac{78000}{2} = 39000$$

  - **چارک اول:** 30000

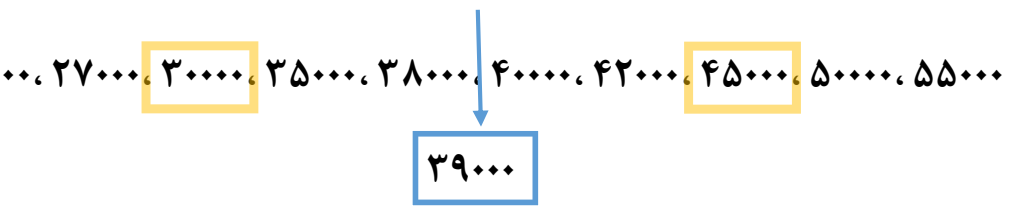
میانه، داده‌ها را به دو بخش تقسیم می‌کند. میانه‌ی بخش اول، همان چارک اول کل داده‌هاست. داده‌های بخش اول:

۲۵۰۰۰، ۲۷۰۰۰، ۳۰۰۰۰، ۳۵۰۰۰، ۳۸۰۰۰

  - **چارک سوم:** 45000

میانه، داده‌ها را به دو بخش تقسیم می‌کند. میانه‌ی بخش دوم، همان چارک سوم کل داده‌هاست. داده‌های بخش دوم:

۴۰۰۰۰، ۴۲۰۰۰، ۴۵۰۰۰، ۵۰۰۰۰، ۵۵۰۰۰



- مجموعه داده‌ی پس‌انداز به صورت مرتب:

۸۰۰۰، ۱۰۰۰۰، ۱۲۰۰۰، ۱۵۰۰۰، ۱۶۰۰۰، ۱۸۰۰۰، ۲۰۰۰۰، ۲۲۰۰۰، ۲۵۰۰۰، ۳۰۰۰۰

  - **میانه:** 17000

عددی که داده‌ها را به دو بخش با تعداد مساوی تقسیم می‌کند، همان میانه است. با توجه به این‌که تعداد داده‌ها زوج است، میانگین دو عدد وسط، به عنوان میانه در نظر گرفته می‌شود:

$$\frac{16000 + 18000}{2} = \frac{34000}{2} = 17000$$

  - **چارک اول:** 12000

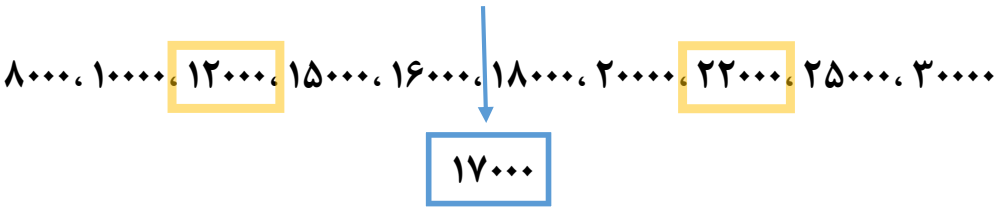
میانه، داده‌ها را به دو بخش تقسیم می‌کند. میانه‌ی بخش اول، همان چارک اول کل داده‌هاست. داده‌های بخش اول:

۸۰۰۰، ۱۰۰۰۰، ۱۲۰۰۰، ۱۵۰۰۰، ۱۶۰۰۰

  - **چارک سوم:** 22000

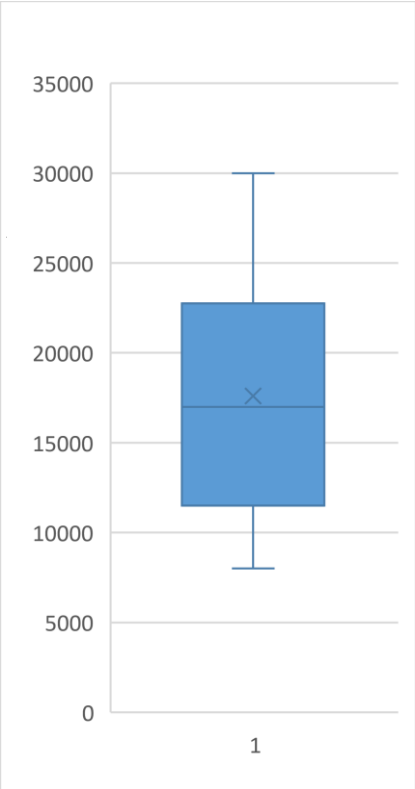
میانه، داده‌ها را به دو بخش تقسیم می‌کند. میانه‌ی بخش دوم، همان چارک سوم کل داده‌هاست. داده‌های بخش دوم:

۱۸۰۰۰، ۲۰۰۰۰، ۲۲۰۰۰، ۲۵۰۰۰، ۳۰۰۰۰



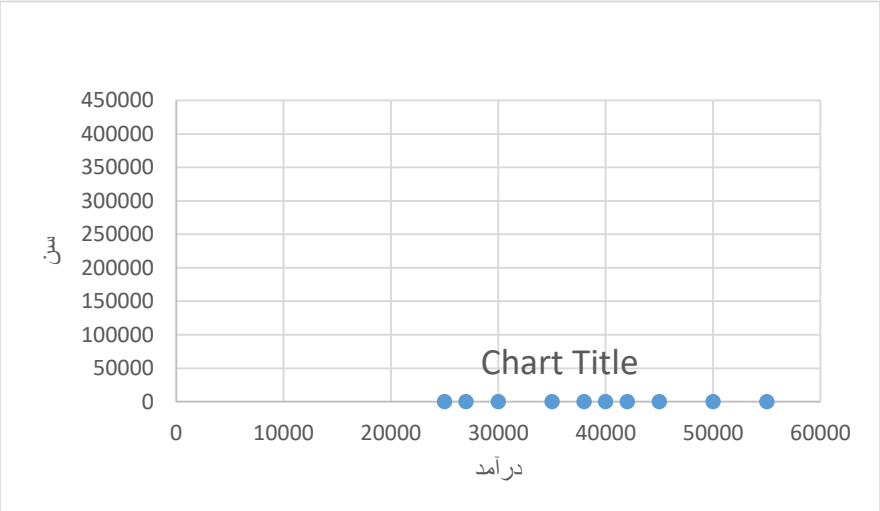
۲.

- نمودار جعبه‌ای برای ویژگی "Savings":



۳.

- نمودار پراکندگی نشان‌دهنده‌ی رابطه‌ی بین "سن" و "درآمد":



۴.

- نمودار Q-Q Plot برای ویژگی "درآمد":

با محاسبه‌ی مقادیر زیر، نمودار رسم شد.

income	rank	percentile	percentile z-score	income z-score
25000	1	0.05	-1.644853627	-1.401416892
27000	2	0.15	-1.036433389	-1.196830485
30000	3	0.25	-0.67448975	-0.889950873
35000	4	0.35	-0.385320466	-0.378484854
38000	5	0.45	-0.125661347	-0.071605243
40000	6	0.55	0.125661347	0.132981165
42000	7	0.65	0.385320466	0.337567573
45000	8	0.75	0.67448975	0.644447184
50000	9	0.85	1.036433389	1.155913203
55000	10	0.95	1.644853627	1.667379222

برای داده‌ها، یک ستون جدید به نام percentile بد ست آمد. و از روی آن، یک مجموعه داده با توزیع نرمال ساخته شد و به صورت نرمالایز شده، در یک ستون جدید اضافه شد.

- $\text{percentile} = \frac{\text{rank}-0.5}{\text{N}} = \frac{\text{rank}-0.5}{10}$

مثلا:

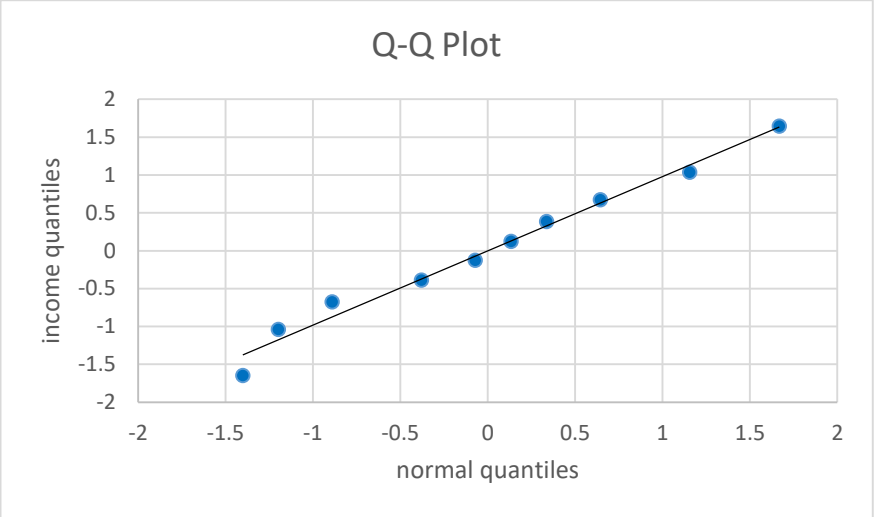
$$\text{percentile } 1 = \frac{1-0.5}{10} = \frac{0.5}{10} = 0.05$$

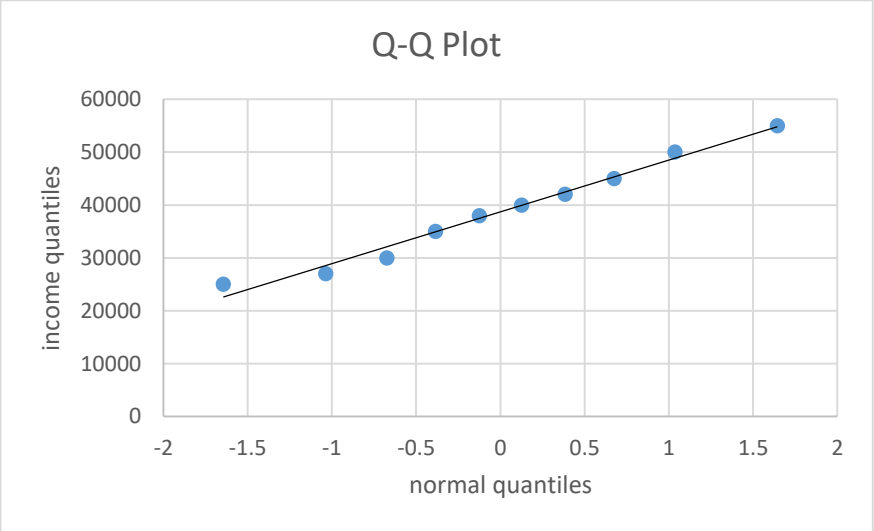
$$\text{percentile } 2 = \frac{2-0.5}{10} = \frac{1.5}{10} = 0.15$$

همچنین، داده‌های مربوط به درآمد به صورت نرمالایز شده در یک ستون جدید اضافه شد.

نمودار Q-Q Plot یکبار برای داده‌های درآمد و یکبار برای داده‌های نرمالایز شده‌ی آن، رسم شد.

با توجه به این‌که داده‌ها در نمودار، منطبق یا نزدیک به خطِ رسم شده هستند، می‌توان نتیجه گرفت که توزیع داده‌ها تقریباً نزدیک به توزیع نرمال است:





۵.

- میانگین مجموعه داده‌ی سن:

$$\overline{\text{age}} = \frac{\sum_{i=1}^{10} \text{age}_i}{N} = \frac{371}{10} = 37.1$$

- میانگین مجموعه داده‌ی درآمد:

$$\overline{\text{income}} = \frac{\sum_{i=1}^{10} \text{income}_i}{N} = \frac{387000}{10} = 38700$$

- میانگین مجموعه داده‌ی پس‌انداز:

$$\overline{\text{savings}} = \frac{\sum_{i=1}^{10} \text{savings}_i}{N} = \frac{176000}{10} = 17600$$

- انحراف از معیار مجموعه داده‌ی سن:

$$\sigma_{age} = \sqrt{\frac{\sum_{i=1}^{10}(age_i - \overline{age})^2}{N}} = 9.230986$$

- انحراف از معیار مجموعه داده‌ی درآمد:

$$\sigma_{income} = \sqrt{\frac{\sum_{i=1}^{10}(income_i - \overline{income})^2}{N}} = 9775.821$$

- انحراف از معیار مجموعه داده‌ی پس‌انداز:

$$\sigma_{savings} = \sqrt{\frac{\sum_{i=1}^{10}(savings_i - \overline{savings})^2}{N}} = 6866.99$$

✓ ضریب همبستگی برای دو ویژگی سن و درآمد:

$$r_{age.income} = \sqrt{\frac{\sum_{i=1}^{10}(age_i - \overline{age})(income_i - \overline{income})}{(N-1)\sigma_{age}\sigma_{income}}} = 0.937372$$

✓ ضریب همبستگی برای دو ویژگی سن و پس‌انداز:

$$r_{age.savings} = \sqrt{\frac{\sum_{i=1}^{10}(age_i - \overline{age})(savings_i - \overline{savings})}{(N-1)\sigma_{age}\sigma_{savings}}} = 0.845571$$

✓ ضریب همبستگی برای دو ویژگی درآمد و پس‌انداز:

$$r_{income.savings} = \sqrt{\frac{\sum_{i=1}^{10}(income_i - \overline{income})(savings_i - \overline{savings})}{(N-1)\sigma_{income}\sigma_{savings}}} = 0.954692$$

سوال ۴:

۱.

- نرمال سازی مجموعه داده‌ی قد با استفاده از تکنیک min-max:

$$v' = \frac{v - min_{height}}{max_{height} - min_{height}}(new\_max_{height} - new\_min_{height}) + new\_min_{height}$$

$$v_1' = \frac{165 - 155}{180 - 155}(1 - 0) + 0 = \frac{10}{25} = 0.4$$
$$v_2' = \frac{170 - 155}{180 - 155}(1 - 0) + 0 = \frac{15}{25} = 0.6$$
$$v_3' = \frac{155 - 155}{180 - 155}(1 - 0) + 0 = \frac{0}{25} = 0$$
$$v_4' = \frac{180 - 155}{180 - 155}(1 - 0) + 0 = \frac{25}{25} = 1$$
$$v_5' = \frac{160 - 155}{180 - 155}(1 - 0) + 0 = \frac{5}{25} = 0.2$$
$$v_6' = \frac{175 - 155}{180 - 155}(1 - 0) + 0 = \frac{20}{25} = 0.8$$

165	0.4
170	0.6
155	0
180	1
160	0.2
175	0.8

- نرمال سازی مجموعه داده‌ی وزن با استفاده از تکنیک min-max:

$$v' = \frac{v - \min_{weight}}{\max_{weight} - \min_{weight}}(new\_max_{weight} - new\_min_{weight}) + new\_min_{weight}$$

$$v_1' = \frac{70 - 45}{90 - 45}(1 - 0) + 0 = \frac{25}{45} = 0.56$$
$$v_2' = \frac{65 - 45}{90 - 45}(1 - 0) + 0 = \frac{20}{45} = 0.44$$
$$v_3' = \frac{45 - 45}{90 - 45}(1 - 0) + 0 = \frac{0}{45} = 0$$
$$v_4' = \frac{90 - 45}{90 - 45}(1 - 0) + 0 = \frac{45}{45} = 1$$
$$v_5' = \frac{50 - 45}{90 - 45}(1 - 0) + 0 = \frac{5}{45} = 0.11$$
$$v_6' = \frac{75 - 45}{90 - 45}(1 - 0) + 0 = \frac{30}{45} = 0.67$$

70	0.6
65	0.4
45	0
90	1
50	0.1
75	0.7

- نرمال سازی مجموعه داده‌ی سن با استفاده از تکنیک min-max:

$$v' = \frac{v - \min_{age}}{\max_{age} - \min_{age}}(new\_max_{age} - new\_min_{age}) + new\_min_{age}$$

$$v_1' = \frac{30 - 25}{40 - 25}(1 - 0) + 0 = \frac{5}{15} = 0.33$$
$$v_2' = \frac{28 - 25}{40 - 25}(1 - 0) + 0 = \frac{3}{15} = 0.2$$

$$v'_3 = \frac{35-25}{40-25}(1-0)+0 = \frac{10}{15} = 0.67$$
$$v'_4 = \frac{40-25}{40-25}(1-0)+0 = \frac{15}{15} = 1$$
$$v'_5 = \frac{25-25}{40-25}(1-0)+0 = \frac{0}{15} = 0$$
$$v'_6 = \frac{32-25}{40-25}(1-0)+0 = \frac{7}{15} = 0.46$$

30	0.3
28	0.2
35	0.7
40	1
25	0
32	0.5

۲.

- نرمال سازی مجموعه داده‌ی قد با استفاده از تکنیک decimal scaling:

$$v' = \frac{v}{10^j}$$

$$v'_1 = \frac{v}{1000} = \frac{165}{1000} = 0.165$$
$$v'_2 = \frac{v}{1000} = \frac{170}{1000} = 0.17$$
$$v'_3 = \frac{v}{1000} = \frac{155}{1000} = 0.155$$
$$v'_4 = \frac{v}{1000} = \frac{180}{1000} = 0.18$$
$$v'_5 = \frac{v}{1000} = \frac{160}{1000} = 0.16$$
$$v'_6 = \frac{v}{1000} = \frac{175}{1000} = 0.175$$

165	0.17
170	0.17
155	0.16
180	0.18
160	0.16
175	0.18

- نرمال سازی مجموعه داده‌ی وزن با استفاده از تکنیک decimal scaling:

$$v' = \frac{v}{10^j}$$

$$v'_1 = \frac{v}{100} = \frac{70}{100} = 0.7$$

$$v'_2 = \frac{v}{100} = \frac{65}{100} = 0.65$$
$$v'_3 = \frac{v}{100} = \frac{45}{100} = 0.45$$
$$v'_4 = \frac{v}{100} = \frac{90}{100} = 0.9$$
$$v'_5 = \frac{v}{100} = \frac{50}{100} = 0.5$$
$$v'_6 = \frac{v}{100} = \frac{75}{100} = 0.75$$

70	0.70
65	0.65
45	0.45
90	0.90
50	0.50
75	0.75

- نرمال سازی مجموعه داده‌ی سن با استفاده از تکنیک decimal scaling:

$$v' = \frac{v}{10^j}$$

$$v'_1 = \frac{v}{100} = \frac{30}{100} = 0.3$$
$$v'_2 = \frac{v}{100} = \frac{28}{100} = 0.28$$
$$v'_3 = \frac{v}{100} = \frac{35}{100} = 0.35$$
$$v'_4 = \frac{v}{100} = \frac{40}{100} = 0.4$$
$$v'_5 = \frac{v}{100} = \frac{25}{100} = 0.25$$
$$v'_6 = \frac{v}{100} = \frac{32}{100} = 0.32$$

30	0.30
28	0.28
35	0.35
40	0.40
25	0.25
32	0.32

.۳

- نرمال سازی مجموعه داده‌ی قد با استفاده از تکنیک z-score:

$$\overline{height} = \frac{\sum_{i=1}^{6} height_i}{N} = \frac{1005}{6} = 167.5$$

$$\sigma_{\text{height}} = \sqrt{\frac{\sum_{i=1}^{6} \left(\text{height}_i - \overline{\text{height}}\right)^2}{N}} = 9.354143$$

$$v' = \frac{v - \overline{\text{height}}}{\sigma_{\text{height}}} = \frac{v - 167.5}{9.354143}$$

$$v'_1 = \frac{v - 167.5}{9.354143} = \frac{165 - 167.5}{9.354143} = \frac{-2.5}{9.354143} = -0.2672$$

$$v'_2 = \frac{v - 167.5}{9.354143} = \frac{170 - 167.5}{9.354143} = \frac{2.5}{9.354143} = 0.2672$$

$$v'_3 = \frac{v - 167.5}{9.354143} = \frac{155 - 167.5}{9.354143} = \frac{-12.5}{9.354143} = -1.3363$$

$$v'_4 = \frac{v - 167.5}{9.354143} = \frac{180 - 167.5}{9.354143} = \frac{12.5}{9.354143} = 1.3363$$

$$v'_5 = \frac{v - 167.5}{9.354143} = \frac{160 - 167.5}{9.354143} = \frac{-7.5}{9.354143} = -0.8017$$

$$v'_6 = \frac{v - 167.5}{9.354143} = \frac{175 - 167.5}{9.354143} = \frac{7.5}{9.354143} = 0.8017$$

165	-0.27
170	0.27
155	-1.34
180	1.34
160	-0.80
175	0.80

- نرمال سازی مجموعه داده‌ی وزن با استفاده از تکنیک z-score:

$$\overline{\text{weight}} = \frac{\sum_{i=1}^{6} \text{weight}_i}{N} = \frac{395}{6} = 65.83$$

$$\sigma_{weight} = \sqrt{\frac{\sum_{i=1}^{6} \left(weight_i - \overline{weight}\right)^2}{N}} = 16.557$$

$$v' = \frac{v - \overline{weight}}{\sigma_{weight}} = \frac{v - 65.83}{16.557}$$

$$v_1' = \frac{v - 65.83}{16.557} = \frac{70 - 65.83}{16.557} = \frac{4.17}{16.557} = 0.2518$$

$$v_2' = \frac{v - 65.83}{16.557} = \frac{65 - 65.83}{16.557} = \frac{0.83}{16.557} = 0.0501$$

$$v_3' = \frac{v - 65.83}{16.557} = \frac{45 - 65.83}{16.557} = \frac{-20.83}{16.557} = -1.258$$

$$v_4' = \frac{v - 65.83}{16.557} = \frac{90 - 65.83}{16.557} = \frac{24.17}{16.557} = 1.459$$

$$v_5' = \frac{v - 65.83}{16.557} = \frac{50 - 65.83}{16.557} = \frac{-15.83}{16.557} = -0.956$$

$$v_6' = \frac{v - 65.83}{16.557} = \frac{75 - 65.83}{16.557} = \frac{9.17}{16.557} = 0.553$$

70	0.25
65	-0.05
45	-1.26
90	1.46
50	-0.96
75	0.55

- نرمال سازی مجموعه داده‌ی سن با استفاده از تکنیک z-score:

$$\overline{age} = \frac{\sum_{i=1}^{6} age_i}{N} = \frac{190}{6} = 31.666$$

$$\sigma_{age} = \sqrt{\frac{\sum_{i=1}^{6}(age_i - \overline{age})^2}{N}} = 5.316$$

$$v' = \frac{v - \overline{age}}{\sigma_{age}} = \frac{v - 31.666}{5.316}$$

$$v'_1 = \frac{v - 31.666}{5.316} = \frac{30 - 31.666}{5.316} = \frac{-1.666}{5.316} = -0.3133$$

$$v'_2 = \frac{v - 31.666}{5.316} = \frac{28 - 31.666}{5.316} = \frac{-3.66}{5.316} = -0.6884$$

$$v'_3 = \frac{v - 31.666}{5.316} = \frac{35 - 31.666}{5.316} = \frac{3.334}{5.316} = 0.6271$$

$$v'_4 = \frac{v - 31.666}{5.316} = \frac{40 - 31.666}{5.316} = \frac{8.334}{5.316} = 1.5677$$

$$v'_5 = \frac{v - 31.666}{5.316} = \frac{25 - 31.666}{5.316} = \frac{-6.666}{5.316} = -1.2539$$

$$v'_6 = \frac{v - 31.666}{5.316} = \frac{32 - 31.666}{5.316} = \frac{0.334}{5.316} = 0.062$$

30	-0.31
28	-0.69
35	0.63
40	1.57
25	-1.25
32	0.06

۴.

رکورد پرس‌وجو به صورت نرمال شده با min-max:

$$v'_{age} = \frac{35-25}{40-25}(1-0)+0 = \frac{10}{15} = 0.66$$
$$v'_{weight} = \frac{80-45}{90-45}(1-0)+0 = \frac{35}{45} = 0.77$$
$$v'_{height} = \frac{175-155}{180-155}(1-0)+0 = \frac{20}{25} = 0.8$$

175	80	35
0.8	0.777778	0.666667

- محاسبه‌ی فاصله‌ی رکورد پرس‌وجو با داده‌های نرمال شده با min-max، با استفاده از سه معیار فاصله:
  - معیار فاصله اقلیدسی:

$$h_e = \sqrt{|height - 0.8|^2 + |weight - 0.77|^2 + |age - 0.66|^2}$$

Record 1:

$$h_e = \sqrt{|0.4-0.8|^2 + |0.6-0.77|^2 + |0.3-0.66|^2}$$
$$= \sqrt{(0.4)^2 + (0.17)^2 + (0.36)^2}$$
$$= \sqrt{0.16 + 0.02 + 0.12} = \sqrt{0.3} = 0.54$$

Record 2:

$$h_e = \sqrt{|0.6-0.8|^2 + |0.4-0.77|^2 + |0.2-0.66|^2}$$
$$= \sqrt{(0.2)^2 + (0.37)^2 + (0.46)^2} = \sqrt{0.04 + 0.13 + 0.21}$$
$$= \sqrt{0.38} = 0.61$$

Record 3:

$$h_e = \sqrt{|0-0.8|^2 + |0-0.77|^2 + |0.7-0.66|^2}$$
$$= \sqrt{(0.8)^2 + (0.77)^2 + (0.04)^2} = \sqrt{0.64 + 0.59 + 0}$$
$$= \sqrt{1.23} = 1.1$$

Record 4:

$$h_e = \sqrt{|1-0.8|^2 + |1-0.77|^2 + |1-0.66|^2}$$
$$= \sqrt{(0.2)^2 + (0.23)^2 + (0.34)^2} = \sqrt{0.04 + 0.05 + 0.11}$$
$$= \sqrt{0.2} = 0.44$$

Record 5:

$$h_e = \sqrt{|0.2-0.8|^2 + |0.1-0.77|^2 + |0-0.66|^2}$$
$$= \sqrt{(0.6)^2 + (0.23)^2 + (0.66)^2} = \sqrt{0.36 + 0.05 + 0.43}$$
$$= \sqrt{0.84} = 0.91$$

Record 6:

$$h_e = \sqrt{|0.8-0.8|^2 + |0.7-0.77|^2 + |0.5-0.66|^2}$$
$$= \sqrt{0 + (0.07)^2 + (0.16)^2} = \sqrt{0 + 0 + 0.02} = \sqrt{0.02}$$
$$= 0.14$$

فاصله‌ی رکورد پرس‌وجو با رکورد شماره ۶ در مجموعه داده‌ها، نسبت به بقیه‌ی فاصله‌ها، کم‌ترین است. پس این رکورد، مشابه‌ترین رکورد با رکورد پرس‌وجو است. رکورد ۶ دارای مقادیر (۳۲ و ۷۵ و ۱۷۵) است.

- معیار فاصله منهتن:

$$h_m = |height - 0.8| + |weight - 0.77| + |age - 0.66|$$

Record 1:

$$h_m = |0.4 - 0.8| + |0.6 - 0.77| + |0.3 - 0.66| = 0.4 + 0.17 + 0.36 = 0.93$$

Record 2:

$$h_m = |0.6 - 0.8| + |0.4 - 0.77| + |0.2 - 0.66| = 0.2 + 0.37 + 0.46 = 1.03$$

Record 3:

$$h_m = |0 - 0.8| + |0 - 0.77| + |0.7 - 0.66| = 0.8 + 0.77 + 0.04 = 1.61$$

Record 4:

$$h_m = |1 - 0.8| + |1 - 0.77| + |1 - 0.66| = 0.2 + 0.23 + 0.34 = 0.77$$

Record 5:

$$h_m = |0.2 - 0.8| + |0.1 - 0.77| + |0 - 0.66| = 0.6 + 0.23 + 0.66 = 1.49$$

Record 6:

$$h_m = |0.8 - 0.8| + |0.7 - 0.77| + |0.5 - 0.66| = 0 + 0.07 + 0.16 = 0.23$$

فاصله‌ی رکورد پرس‌وجو با رکورد شماره ۶ در مجموعه داده‌ها، نسبت به بقیه‌ی فاصله‌ها، کم‌ترین است. پس این رکورد، مشابه‌ترین رکورد با رکورد پرس‌وجو است. رکورد ۶ دارای مقادیر (۳۲ و ۷۵ و ۱۷۵) است.

- معیار فاصله supremum:

$$h_s = \max\ (|\text{height} - 0.8| \,.\, |\text{weight} - 0.77| \,.\, |\text{age} - 0.66|)$$

Record 1:

$$h_s = \max\ (|0.4 - 0.8| \,.\, |0.6 - 0.77| \,.\, |0.3 - 0.66|)$$
$$= \max\ (0.4 \,.\, 0.17 \,.\, 0.36) = 0.4$$

Record 2:

$$h_s = \max\ (|0.6 - 0.8| \,.\, |0.4 - 0.77| \,.\, |0.2 - 0.66|)$$
$$= \max\ (0.2 \,.\, 0.37 \,.0.46) = 0.46$$

Record 3:

$$h_s = \max\ (|0 - 0.8| \,.\, |0 - 0.77| \,.\, |0.7 - 0.66|)$$
$$= \max\ (0.8 \,.\, 0.77 \,.\, 0.04) = 0.8$$

Record 4:

$$h_s = \max\ (|1 - 0.8| \,.\, |1 - 0.77| \,.\, |1 - 0.66|)$$
$$= \max\ (0.2 \,.\, 0.23 \,.\, 0.34) = 0.34$$

Record 5:

$$h_s = \max\ (|0.2 - 0.8| \,.\, |0.1 - 0.77| \,.\, |0 - 0.66|)$$
$$= \max\ (0.6 \,.\, 0.23 \,.\ 0.66) = 0.66$$

Record 6:

$$h_s = \max\ (|0.8 - 0.8| \,.\, |0.7 - 0.77| \,.\, |0.5 - 0.66|)$$
$$= \max\ (0 \,.\, 0.07 \,.0.16) = 0.16$$

فاصله‌ی رکورد پرس‌وجو با رکورد شماره ۶ در مجموعه داده‌ها، نسبت به بقیه‌ی فاصله‌ها، کم‌ترین است. پس این رکورد، مشابه‌ترین رکورد با رکورد پرس‌وجو است. رکورد ۶ دارای مقادیر (۳۲ و ۷۵ و ۱۷۵) است.

۵.

رکورد پرس‌وجو به صورت نرمال شده با decimal scaling:

$$v'_{age} = \frac{v}{100} = \frac{35}{100} = 0.35$$

$$v'_{weight} = \frac{v}{100} = \frac{80}{100} = 0.8$$

$$v'_{height} = \frac{v}{1000} = \frac{175}{1000} = 0.175$$

175	80	35
0.175	0.8	0.35

- محاسبه‌ی فاصله‌ی رکورد پرس‌وجو با داده‌های نرمال شده با decimal scaling، با استفاده از سه معیار فاصله:
  - معیار فاصله اقلیدسی:

$$h_e = \sqrt{|height - 0.175|^2 + |weight - 0.8|^2 + |age - 0.35|^2}$$

Record 1:

$$h_e = \sqrt{|0.17 - 0.175|^2 + |0.7 - 0.8|^2 + |0.3 - 0.35|^2}$$
$$= \sqrt{(0.005)^2 + (0.1)^2 + (0.05)^2} = \sqrt{0.01} = 0.1$$

Record 2:

$$h_e = \sqrt{|0.17 - 0.175|^2 + |0.65 - 0.8|^2 + |0.28 - 0.35|^2}$$
$$= \sqrt{(0.005)^2 + (0.15)^2 + (0.7)^2} = \sqrt{0 + 0.02 + 0.49}$$
$$= \sqrt{0.51} = 0.71$$

Record 3:

$$h_e = \sqrt{|0.16 - 0.175|^2 + |0.45 - 0.8|^2 + |0.35 - 0.35|^2}$$
$$= \sqrt{(0.015)^2 + (0.35)^2 + (0)^2} = \sqrt{0.12} = 0.34$$

Record 4:

$$h_e = \sqrt{|0.18 - 0.175|^2 + |0.9 - 0.8|^2 + |0.4 - 0.35|^2}$$
$$= \sqrt{(0.005)^2 + (0.1)^2 + (0.05)^2} = \sqrt{0.01} = 0.1$$

Record 5:

$$h_e = \sqrt{|0.16 - 0.175|^2 + |0.5 - 0.8|^2 + |0.25 - 0.35|^2}$$
$$= \sqrt{(0.015)^2 + (0.3)^2 + (0.1)^2} = \sqrt{0.09 + 0.01}$$
$$= \sqrt{0.1} = 0.31$$

Record 6:

$$h_e = \sqrt{|0.18 - 0.175|^2 + |0.75 - 0.8|^2 + |0.32 - 0.35|^2}$$
$$= \sqrt{(0.005)^2 + (0.05)^2 + (0.03)^2} = \sqrt{0.002 + 0.0009}$$
$$= \sqrt{0.0029} = 0.05$$

فاصله‌ی رکورد پرس‌وجو با رکورد شماره ۶ در مجموعه داده‌ها، نسبت به بقیه‌ی فاصله‌ها، کم‌ترین است. پس این رکورد، مشابه‌ترین رکورد با رکورد پرس‌وجو است. رکورد ۶ دارای مقادیر (۳۲ و ۷۵ و ۱۷۵) است.

- معیار فاصله منهتن:

$$h_m = |height - 0.175| + |weight - 0.8| + |age - 0.35|$$

Record 1:

$$h_m = |0.17 - 0.175| + |0.7 - 0.8| + |0.3 - 0.35|$$
$$= 0.005 + 0.1 + 0.05 = 0.155$$

Record 2:

$$h_m = |0.17 - 0.175| + |0.65 - 0.8| + |0.28 - 0.35|$$
$$= 0.005 + 0.15 + 0.07 = 0.225$$

Record 3:

$$h_m = |0.16 - 0.175| + |0.45 - 0.8| + |0.35 - 0.35|$$
$$= 0.015 + 0.35 + 0 = 0.365$$

Record 4:

$$h_m = |0.18 - 0.175| + |0.9 - 0.8| + |0.4 - 0.35|$$
$$= 0.005 + 0.1 + 0.05 = 0.155$$

Record 5:

$$h_m = |0.16 - 0.175| + |0.5 - 0.8| + |0.25 - 0.35|$$
$$= 0.015 + 0.3 + 0.1 = 0.415$$

Record 6:

$$h_m = |0.18 - 0.175| + |0.75 - 0.8| + |0.32 - 0.35|$$
$$= 0.005 + 0.05 + 0.03 = 0.085$$

فاصله‌ی رکورد پرس‌وجو با رکورد شماره ۶ در مجموعه داده‌ها، نسبت به بقیه‌ی فاصله‌ها، کم‌ترین است. پس این رکورد، مشابه‌ترین رکورد با رکورد پرس‌وجو است. رکورد ۶ دارای مقادیر (۳۲ و ۷۵ و ۱۷۵) است.

- معیار فاصله supremum:

$$h_s = \max\ (|height - 0.175| \,.\, |weight - 0.8| \,.\, |age - 0.35|)$$

Record 1:

$$h_s = \max\ (|0.17 - 0.175| \,.\, |0.7 - 0.8| \,.\, |0.3 - 0.35|)$$
$$= \max\ (0.005 \,.\, 0.1 \,.\, 0.05) = 0.1$$

Record 2:

$$h_s = \max\ (|0.17 - 0.175| \,.\, |0.65 - 0.8| \,.\, |0.28 - 0.35|)$$
$$= \max\ (0.005 \,.\, 0.15 \,.\, 0.07) = 0.225$$

Record 3:

$$h_s = \max\ (|0.16 - 0.175| \,.\, |0.45 - 0.8| \,.\, |0.35 - 0.35|)$$
$$= \max\ (0.015 \,.\, 0.35 \,.\, 0) = 0.35$$

Record 4:

$$h_s = \max\ (|0.18 - 0.175| \,.\, |0.9 - 0.8| \,.\, |0.4 - 0.35|)$$
$$= \max\ (0.005 \,.\, 0.1 \,.\, 0.05) = 0.1$$

Record 5:

$$h_s = \max\ (|0.16 - 0.175| \,.\, |0.5 - 0.8| \,.\, |0.25 - 0.35|)$$
$$= \max\ (0.015 \,.\, 0.3 \,.\, 0.1) = 0.3$$

Record 6:

$$h_s = \max\ (|0.18 - 0.175| \,.\, |0.75 - 0.8| \,.\, |0.32 - 0.35|)$$
$$= \max\ (0.005 \,.\, 0.05 \,.\, 0.03) = 0.05$$

فاصله‌ی رکورد پرس‌وجو با رکورد شماره ۶ در مجموعه داده‌ها، نسبت به بقیه‌ی فاصله‌ها، کم‌ترین است. پس این رکورد، مشابه‌ترین رکورد با رکورد پرس‌وجو است. رکورد ۶ دارای مقادیر (۳۲ و ۷۵ و ۱۷۵) است.

۶.

رکورد پرس‌وجو به صورت نرمال شده با z-score:

$$v'_{age} = \frac{v - \overline{age}}{\sigma_{age}} = \frac{v - 31.666}{5.316} = \frac{35 - 31.666}{5.316} = \frac{3.334}{5.316} = 0.62$$

$$v'_{weight} = \frac{v - \overline{weight}}{\sigma_{weight}} = \frac{v - 65.83}{16.557} = \frac{80 - 65.83}{16.557} = \frac{14.17}{16.557} = 0.85$$

$$v'_{height} = \frac{v - \overline{height}}{\sigma_{height}} = \frac{v - 167.5}{9.354143} = \frac{175 - 167.5}{9.354143} = \frac{7.5}{9.354143} = 0.8$$

175	80	35
0.801784	0.855579	0.626962

- محاسبه‌ی فاصله‌ی رکورد پرس‌وجو با داده‌های نرمال شده با z-score، با استفاده از سه معیار فاصله:
  - معیار فاصله اقلیدسی:

$$h_e = \sqrt{|\text{height} - 0.8|^2 + |\text{weight} - 0.85|^2 + |\text{age} - 0.62|^2}$$

Record 1:

$$h_e = \sqrt{|-0.27 - 0.8|^2 + |0.25 - 0.85|^2 + |-0.31 - 0.62|^2}$$
$$= \sqrt{(1.07)^2 + (0.6)^2 + (0.93)^2} = \sqrt{1.14 + 0.36 + 0.86}$$
$$= \sqrt{2.36} = 1.53$$

Record 2:

$$h_e = \sqrt{|0.27 - 0.8|^2 + |-0.05 - 0.85|^2 + |-0.69 - 0.62|^2}$$
$$= \sqrt{(0.53)^2 + (0.9)^2 + (1.31)^2} = \sqrt{0.28 + 0.81 + 1.71}$$
$$= \sqrt{2.8} = 1.67$$

Record 3:

$$h_e = \sqrt{|-1.34 - 0.8|^2 + |-1.26 - 0.85|^2 + |0.63 - 0.62|^2}$$
$$= \sqrt{(2.14)^2 + (2.11)^2 + (0.01)^2} = \sqrt{4.57 + 4.45 + 0}$$
$$= \sqrt{9.02} = 3$$

Record 4:

$$h_e = \sqrt{|1.34 - 0.8|^2 + |1.46 - 0.85|^2 + |1.57 - 0.62|^2}$$
$$= \sqrt{(0.54)^2 + (0.61)^2 + (0.95)^2} = \sqrt{0.29 + 0.37 + 0.9}$$
$$= \sqrt{1.56} = 1.24$$

Record 5:

$$h_e = \sqrt{|-0.8-0.8|^2 + |-0.96-0.85|^2 + |-1.25-0.62|^2}$$
$$= \sqrt{(1.6)^2 + (1.81)^2 + (1.87)^2}$$
$$= \sqrt{2.56 + 3.27 + 3.49} = \sqrt{9.32} = 3.05$$

Record 6:

$$h_e = \sqrt{|0.8-0.8|^2 + |0.55-0.85|^2 + |0.06-0.62|^2}$$
$$= \sqrt{(0)^2 + (0.3)^2 + (0.56)^2} = \sqrt{0 + 0.09 + 0.31}$$
$$= \sqrt{0.4} = 0.63$$

فاصله‌ی رکورد پرس‌وجو با رکورد شماره ۶ در مجموعه داده‌ها، نسبت به بقیه‌ی فاصله‌ها، کم‌ترین است. پس این رکورد، مشابه‌ترین رکورد با رکورد پرس‌وجو است. رکورد ۶ دارای مقادیر (۳۲ و ۷۵ و ۱۷۵) است.

- معیار فاصله منهتن:

$$h_m = |height - 0.8| + |weight - 0.85| + |age - 0.62|$$

Record 1:

$$h_m = |-0.27-0.8| + |0.25-0.85| + |-0.31-0.62|$$
$$= 1.07 + 0.6 + 0.93 = 2.6$$

Record 2:

$$h_m = |0.27-0.8| + |-0.05-0.85| + |-0.69-0.62|$$
$$= 0.53 + 0.9 + 1.31 = 2.74$$

Record 3:

$$h_m = |-1.34 - 0.8| + |-1.26 - 0.85| + |0.63 - 0.62|$$
$$= 2.14 + 2.11 + 0.01 = 4.26$$

Record 4:

$$h_m = |1.34 - 0.8| + |1.46 - 0.85| + |1.57 - 0.62|$$
$$= 0.54 + 0.61 + 0.95 = 2.1$$

Record 5:

$$h_m = |-0.8 - 0.8| + |-0.96 - 0.85| + |-1.25 - 0.62|$$
$$= 1.6 + 1.81 + 1.87 = 5.28$$

Record 6:

$$h_m = |0.8 - 0.8| + |0.55 - 0.85| + |0.06 - 0.62| = 0 + 0.3 + 0.56$$
$$= 0.86$$

فاصله‌ی رکورد پرس‌وجو با رکورد شماره ۶ در مجموعه داده‌ها، نسبت به بقیه‌ی فاصله‌ها، کم‌ترین است. پس این رکورد، مشابه‌ترین رکورد با رکورد پرس‌وجو است. رکورد ۶ دارای مقادیر (۳۲ و ۷۵ و ۱۷۵) است.

- معیار فاصله supremum:

$$h_s = \max\ (|\text{height} - 0.8| \,.\, |\text{weight} - 0.85| \,.\, |\text{age} - 0.62|)$$

Record 1:

$$h_s = \max\ (|-0.27 - 0.8| \,.\, |0.25 - 0.85| \,.\, |-031 - 0.62|)$$
$$= \max\ (1.7 \,.\, 0.6 \,.\, 0.93) = 1.7$$

Record 2:

$$h_s = \max\ (|0.27 - 0.8| \,.\, |-0.05 - 0.85| \,.\, |-0.69 - 0.62|)$$
$$= \max\ (0.19 \,.\, 0.9 \,.\, 1.31) = 1.31$$

Record 3:

$$h_s = \max\ (|-1.34 - 0.8| \,.\, |-1.26 - 0.85| \,.\, |0.63 - 0.62|)$$
$$= \max\ (2.14 \,.\, 2.11 \,.\, 0.01) = 2.14$$

Record 4:

$$h_s = \max\ (|1.34 - 0.8| \,.\, |1.46 - 0.85| \,.\, |1.57 - 0.62|)$$
$$= \max\ (0.54 \,.\, 0.61 \,.\, 0.95) = 0.95$$

Record 5:

$$h_s = \max\ (|-0.8 - 0.8| \,.\, |-0.96 - 0.85| \,.\, |-1.25 - 0.62|)$$
$$= \max\ (1.6 \,.\, 1.81 \,.\, 1.87) = 1.87$$

Record 6:

$$h_s = \max\ (|0.8 - 0.8| \,.\, |0.55 - 0.85| \,.\, |0.06 - 0.62|)$$
$$= \max\ (0 \,.\, 0.3 \,.\, 0.56) = 0.56$$

فاصله‌ی رکورد پرس‌وجو با رکورد شماره ۶ در مجموعه داده‌ها، نسبت به بقیه‌ی فاصله‌ها، کم‌ترین است. پس این رکورد، مشابه‌ترین رکورد با رکورد پرس‌وجو است. رکورد ۶ دارای مقادیر (۳۲ و ۷۵ و ۱۷۵) است.