داده کاوی - تمرین سری اول

دانشگاه صنعتی اصفهان

دانشکده مهندسی برق و کامپیوتر

عنوان: تکلیف اول درس داده کاوی

نام و نام خانوادگی: نیلوفر سعیدی

شماره دانشجویی: ۹۸۲۲۹۶۳

سوال ۱. فرض کنید داده هایی در مورد مراجعان یک بیمارستان در دسترس است. این داده ها می تواند شامل سن، جنسیت، سابقه بیماری قبلی، شغل، قد، وزن و ... باشد. با درنظرگیری داده های موجود چهار نمونه مسئله با استفاده از تسک های داده کاوی مانند پیش بینی، دسته بندی و ... معرفی و توصیف نمایید.

1. Classification: group the jobs into three classes of healthy, mid, and harmful depending on the health condition of the customers being in those jobs for more than 5 years.
2. Regression: how does weight change with age normally in healthiest customers?
3. is there a high correlation between BMI ($\frac{weight}{height^2}$) and heart condition? Prediction: how much is a person with a specific BMI likely to face serious heart condition within 5 years?
4. Regression: predict the healthy BMI in all ranges of age gender-specifically.

سوال ۲. ـصفات زیر را در دسته های ارائه شده طبقه بندی کنید. در صورت ابهام با توضیح دلیل انتخاب خود را بیان کنید.

- سن بر حسب سال

Not Binary - Discrete - Ratio

- روشنایی که با نورسنج اندازه گیری می شود

Not Binary - Continuous - Ratio

- روشنایی که با نظر افراد بیان می شود

Not Binary(bright, very bright, slightly bright, dim, dark, ...) - Discrete - Ordinal

- زاویه اندازه گیری شده با وسیله اندازه گیر ی (نقاله و ...)

Not Binary - Continuous - Interval(can be negative)

- مدال های اهدایی در مسابقات المپیک

Not Binary - Discrete - Ordinal

- ارتفاع از سطح دریا

Not Binary - Continuous - Ratio

- تعداد بیماران یک بیمارستان

Not Binary - Discrete - Ratio

- شماره ISBN

Not Binary - Discrete - Nominal

سوال ۳. احتمال این را پیدا کنید که در انتخاب چهار عدد از ۱ تا ۱۰۰ به طور تصادفی و بدون جدول زیر مقادیر ثبت شده برای قیمت کاالهای استوک وارداتی یک شرکت را نشان می دهد.

لطفا مقادیر زیر را محاسبه کنید:	۱۰	۷	۲۰	۱۲	۷۵	۱۵	۹	۱۸	۴	۱۲	۸	۱۴

۱. میانگین: ۱۷

۲. میانه: ۱۲

۳. مد: ۱۲

۴. انحراف معیار: $\sigma^2 = 325$، $\sigma = 18.02$

۵. zscore شاخص: $z = \frac{x-\mu}{\sigma}$.

For each given record, zscore is as below: (value:zscore)
(10: -0.38), (7: -0.55), (20: 0.16), (12: -0.27), (75: 3.21), (15: -0.11), (9: -0.44), (18: 0.05), (4: -0.72), (8: -0.49), (14: -0.16)

سوال ۴.

الف. تمرین ۲

2.2 Suppose that the data for analysis includes the attribute *age*. The *age* values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

(a) What is the *mean* of the data? What is the *median*?

(b) What is the *mode* of the data? Comment on the data's modality (i.e., bimodal, trimodal, etc.).

(c) What is the *midrange* of the data?

(d) Can you find (roughly) the first quartile (Q_1) and the third quartile (Q_3) of the data?

(e) Give the *five-number summary* of the data.

(f) Show a *boxplot* of the data.

(g) How is a *quantile–quantile plot* different from a *quantile plot*?

a) mean $= 809/27 = 30$

median $= 25$

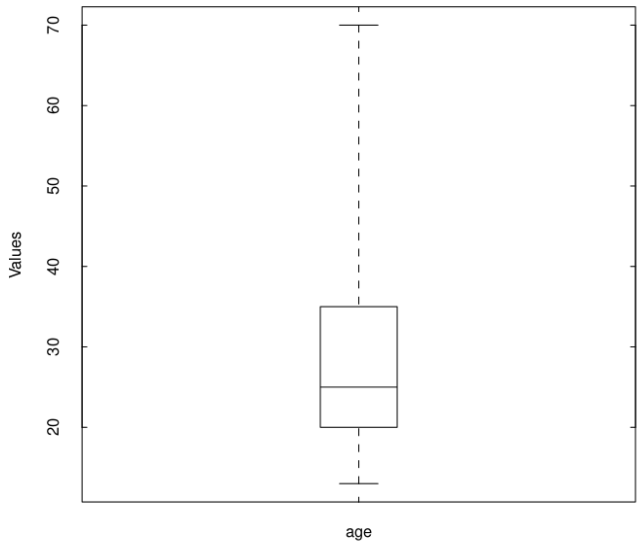
b) 25 and 35 have the same highest frequency and are both modes. So the data is bimodal.

c) midrange $= \frac{max+min}{2} = \frac{70+13}{2} = 41.5$

d) $Q_1 = 20, Q_3 = 35$

e) what?

f)



g) A quantile plot is a graphical method used to show the approximate percentage of values below or equal to the independent variable in a univariate distribution. Thus, it displays quantile information for all the data, where the values measured for the independent variable are plotted against their corresponding quantile. A quantile-quantile

plot however, graphs the quantiles of one univariate distribution against the corresponding quantiles of another univariate distribution. Both axes display the range of values measured for their corresponding distribution, and points are plotted that correspond to the quantile values of the two distributions. A line (y = x) can be added to the graph along with points representing where the first, second and third quantiles lie, in order to increase the graph's informational value. Points that lie above such a line indicate a correspondingly higher value for the distribution plotted on the y-axis, than for the distribution plotted on the x-axis at the same quantile. The opposite effect is true for points lying below this line.

ب. تمرین ۶

2.6 Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8):

(a) Compute the *Euclidean distance* between the two objects.
(b) Compute the *Manhattan distance* between the two objects.
(c) Compute the *Minkowski distance* between the two objects, using $q = 3$.
(d) Compute the *supremum distance* between the two objects.

a) $\sqrt{(22-20)^2 + (1-0)^2 + (42-36)^2 + (10-8)^2} = \sqrt{45} = 6.71$

b) $|22-20|^2 + |1-0|^2 + |42-36|^2 + |10-8|^2 = 11$

c) $\sqrt[3]{(22-20)^3 + (1-0)^3 + (42-36)^3 + (10-8)^3} = \sqrt[3]{233} = 6.15$

d) $|24-36| = 6$