تمرین سری اول داده‌کاوی **متین جعفرآقایی-۹۶۳۱۹۰۳**

سوال ۱-

الف) پیوسته، کمی – نسبت

ب) پیوسته، کمی – بازه

ج) پیوسته، کمی – نسبت

د) گسسته، کیفی – ترتیبی

ح) پیوسته، کمی – بازه

ت) گسسته، کیفی – ترتیبی

سوال ۲-

الف) نویز هیچ‌گاه مطلوب نیست و تنها داده‌های اصلی را خراب می‌کند. outlierها می‌توانند داده‌های مهمی باشند و گاه هدف اصلی داده‌کاوی هستند. بنابراین outlierها ممکن است مطلوب باشند اما نویز طبق تعریف نامطلوب است.

ب) بله، نویز داده‌ها را تصادفی و غیرعادی می‌کند. بنابراین این امکان وجود دارد که نویز به صورت outlier ظاهر شود.

ج) خیر، نویز می‌تواند مانند داده معمولی نیز ظاهر شود.

د) خیر، outlierها می‌توانند داده‌های مفیدی باشند که تنها بنظر می‌رسد به مجموعه داده‌ها تعلق ندارند و لزوما نویز نیستند.

سوال ۳-

الف)

$x = (0, -1, 0, 1)$
$y = (1, 0, -1, 0)$

Euclidean:

$$d(x,y) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2} = \sqrt{(0-1)^2 + (-1-0)^2 + \big(0-(-1)\big)^2 + (1-0)^2} = \sqrt{4} = 2$$

Correlation:

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k = \frac{1}{4}(0 - 1 + 0 + 1) = 0$$
$$\bar{y} = \frac{1}{n}\sum_{k=1}^{n} y_k = \frac{1}{4}(1 + 0 - 1 + 0) = 0$$
$$s_x = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})^2} = \sqrt{\frac{1}{3}((0-0)^2 + (-1-0)^2 + (0-0)^2 + (1-0)^2} = \sqrt{\frac{2}{3}}$$
$$s_y = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(y_k - \bar{y})^2} = \sqrt{\frac{1}{3}((1-0)^2 + (0-0)^2 + (-1-0)^2 + (0-0)^2)} = \sqrt{\frac{2}{3}}$$
$$s_{xy} = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})(y_k - \bar{y})$$
$$= \frac{1}{3}\big((0-0)(1-0) + (-1-0)(0-0) + (0-0)(-1-0) + (1-0)(0-0)\big) = 0$$
$$corr(x,y) = \frac{s_{xy}}{s_x s_y} = \frac{0}{\sqrt{\frac{2}{3}} \times \sqrt{\frac{2}{3}}} = 0$$

Cosine:

$$< x, y > = 0 * 1 + (-1) * 0 + 0 * (-1) + 1 * 0 = 0$$
$$\big||x|\big| = \sqrt{0 * 0 + (-1) * (-1) + 0 * 0 + 1 * 1} = \sqrt{2}$$
$$\big||y|\big| = \sqrt{1 * 1 + 0 * 0 + (-1) * (-1) + 0 * 0} = \sqrt{2}$$
$$\cos(x,y) = \frac{< x, y >}{\big||x|\big|\big||y|\big|} = \frac{0}{\sqrt{2} \times \sqrt{2}} = 0$$

ب)

$$x = (2, -1, 0, 2, 0, -3)$$
$$y = (-1, 1, -1, 0, 0, -1)$$

Euclidean:

$$d(x,y) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2}$$
$$= \sqrt{(2-(-1))^2 + (-1-1)^2 + \big(0-(-1)\big)^2 + (2-0)^2 + (0-0)^2 + (-3-(-1))^2}$$
$$= \sqrt{9 + 4 + 1 + 4 + 0 + 4} = \sqrt{22}$$

Correlation:

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k = \frac{1}{6}(2 - 1 + 0 + 2 + 0 - 3) = 0$$

$$\bar{y} = \frac{1}{n}\sum_{k=1}^{n} y_k = \frac{1}{6}(-1 + 1 - 1 + 0 + 0 - 1) = 0$$

$$s_x = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})^2}$$

$$= \sqrt{\frac{1}{5}((2-0)^2 + (-1-0)^2 + (0-0)^2 + (2-0)^2 + (0-0)^2 + (-3-0)^2} = \frac{3\sqrt{10}}{5}$$

$$s_y = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(y_k - \bar{y})^2}$$

$$= \sqrt{\frac{1}{5}((-1-0)^2 + (1-0)^2 + (-1-0)^2 + (0-0)^2 + (0-0)^2 + (-1-0)^2)}$$

$$= \frac{2\sqrt{5}}{5}$$

$$s_{xy} = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})(y_k - \bar{y})$$

$$= \frac{1}{5}\big((2-0)(-1-0) + (-1-0)(1-0) + (0-0)(-1-0) + (2-0)(0-0)$$

$$+ (0-0)(0-0) + (-3-0)(-1-0)\big) = \frac{1}{5}(-2 - 1 + 0 + 0 + 0 + 3) = 0$$

$$corr(x, y) = \frac{s_{xy}}{s_x s_y} = \frac{0}{\frac{3\sqrt{10}}{5} \times \frac{2\sqrt{5}}{5}} = 0$$

Cosine:

$$< x, y > = 2 * (-1) + (-1) * 1 + 0 * (-1) + 2 * 0 + 0 * 0 + (-3) * (-1) = -2 - 1 + 3 = 0$$

$$||x|| = \sqrt{2 * 2 + (-1) * (-1) + 0 * 0 + 2 * 2 + 0 * 0 + (-3) * (-3)} = \sqrt{4 + 1 + 0 + 4 + 0 + 9}$$

$$= 3\sqrt{2}$$

$$||y|| = \sqrt{(-1) * (-1) + 1 * 1 + (-1) * (-1) + 0 * 0 + 0 * 0 + (-1) * (-1)} = 2$$

$$\cos(x, y) = \frac{< x, y >}{||x||||y||} = \frac{0}{3\sqrt{2} \times 2} = 0$$

$x = (1,1,0,1,0,1)$
$y = (1,1,1,0,0,1)$

Jaccard:

$$J = \frac{f_{11}}{f_{01} + f_{01} + f_{11}} = \frac{3}{1+1+3} = \frac{3}{5}$$

Correlation:

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k = \frac{1}{6}(1+1+0+1+0+1) = \frac{2}{3}$$

$$\bar{y} = \frac{1}{n}\sum_{k=1}^{n} y_k = \frac{1}{6}(1+1+1+0+0+1) = \frac{2}{3}$$

$$s_x = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})^2}$$

$$= \sqrt{\frac{1}{5}\left(\left(1-\frac{2}{3}\right)^2 + \left(1-\frac{2}{3}\right)^2 + \left(0-\frac{2}{3}\right)^2 + \left(1-\frac{2}{3}\right)^2 + \left(0-\frac{2}{3}\right)^2 + \left(1-\frac{2}{3}\right)^2\right)}$$

$$= \sqrt{\frac{1}{5}\left(\frac{1}{9}+\frac{1}{9}+\frac{4}{9}+\frac{1}{9}+\frac{4}{9}+\frac{1}{9}\right)} = \frac{2\sqrt{15}}{15}$$

$$s_y = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(y_k - \bar{y})^2}$$

$$= \sqrt{\frac{1}{5}\left(\left(1-\frac{2}{3}\right)^2 + \left(1-\frac{2}{3}\right)^2 + \left(1-\frac{2}{3}\right)^2 + \left(0-\frac{2}{3}\right)^2 + \left(0-\frac{2}{3}\right)^2 + \left(1-\frac{2}{3}\right)^2\right)} =$$

$$= \sqrt{\frac{1}{5}\left(\frac{1}{9}+\frac{1}{9}+\frac{1}{9}+\frac{4}{9}+\frac{4}{9}+\frac{1}{9}\right)} = \frac{2\sqrt{15}}{15}$$

$$s_{xy} = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})(y_k - \bar{y})$$

$$= \frac{1}{5}\left(\left(1-\frac{2}{3}\right)\left(1-\frac{2}{3}\right) + \left(1-\frac{2}{3}\right)\left(1-\frac{2}{3}\right) + \left(0-\frac{2}{3}\right)\left(1-\frac{2}{3}\right) + \left(1-\frac{2}{3}\right)\left(0-\frac{2}{3}\right)\right.$$

$$\left. + \left(0-\frac{2}{3}\right)\left(0-\frac{2}{3}\right) + \left(1-\frac{2}{3}\right)\left(1-\frac{2}{3}\right)\right) = \frac{1}{5}\left(\frac{1}{9}+\frac{1}{9}-\frac{2}{9}-\frac{2}{9}+\frac{4}{9}+\frac{1}{9}\right) = \frac{1}{15}$$

$$corr(x,y) = \frac{s_{xy}}{s_x s_y} = \frac{\frac{1}{15}}{\frac{2\sqrt{15}}{15} \times \frac{2\sqrt{15}}{15}} = \frac{1}{4}$$

Cosine:

$$< x, y > = 1 * 1 + 1 * 1 + 0 * 1 + 1 * 0 + 0 * 0 + 1 * 1 = 3$$
$$||x|| = \sqrt{1 * 1 + 1 * 1 + 0 * 0 + 1 * 1 + 0 * 0 + 1 * 1} = 2$$
$$||y|| = \sqrt{1 * 1 + 1 * 1 + 1 * 1 + 0 * 0 + 0 * 0 + 1 * 1} = 2$$
$$\cos(x, y) = \frac{< x, y >}{||x|| ||y||} = \frac{3}{2 \times 2} = \frac{3}{4}$$

سوال ۴-

الف)

راه حل اول: دور داده‌هایی که برخی ویژگی‌های آن‌ها حذف شده است.

راه حل دوم: تخمین مقدار ویژگی‌های حذف شده با استفاده از داده‌های مجاور

ب)

راه حل اول: کاهش ابعاد با استفاده از روش‌هایی مانند PCA

راه حل دوم: حذف ویژگی‌های زائد و غیر مرتبط

ج)

راه حل اول: جمع‌آوری داده‌های بیشتر برای متعادل کردن تعداد داده‌های با برچسب متفاوت

راه حل دوم: انتخاب نمونه‌ای از داده‌های موجود که از همه‌ی برچسب‌ها تعداد کافی داشته باشد.

د)

راه حل اول: ساده‌تر کردن مدل آموزشی با کاهش تعداد داده‌های یادگیری

راه حل دوم: استفاده از روش‌های regularization مانند Tikhonov

ه)

راه حل اول: استفاده از تعداد داده‌های بیشتر برای آموزش و کاهش خطای مدل آموزشی بر روی آن‌ها

راه حل دوم: پیچیده‌تر کردن مدل آموزشی مثلا با افزایش عمق آن

سوال ۵-

الف) بله، اگر تعداد داده‌های outlier زیاد باشد، می‌تواند باعث منحرف شدن رابطه خطی بدست آمده در رگرسیون خطی شود.

ب) از جمع مربعات فاصله نقاط از خط پیشبینی شده برای محاسبه خطا استفاده می‌شود. از آنجاییکه این تابع مشتق پذیر است و مینیمم محلی ندارد، به عنوان معیار اصلی اندازه‌گیری خطا از آن استفاده می‌شود.

ج)

$$X = \begin{bmatrix} 1 & 1.58 \\ 1 & 1.60 \\ 1 & 1.62 \\ 1 & 1.65 \\ 1 & 1.68 \\ 1 & 1.70 \\ 1 & 1.74 \\ 1 & 1.75 \\ 1 & 1.77 \\ 1 & 1.80 \end{bmatrix}, \qquad y = \begin{bmatrix} 57.5 \\ 58.2 \\ 59.5 \\ 62.1 \\ 63.4 \\ 64.5 \\ 66.2 \\ 67.7 \\ 69.4 \\ 71.3 \end{bmatrix}, \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

$$X^T X\beta = X^t y \rightarrow \begin{bmatrix} 10 & 16.89 \\ 16.89 & 28.58 \end{bmatrix}\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} 639.8 \\ 1083.83 \end{bmatrix} \rightarrow \begin{cases} 10\beta_0 + 16.89\beta_1 = 639.8 \\ 16.89\beta_0 + 28.58\beta_1 = 1083.83 \end{cases}$$
$$\rightarrow \begin{cases} \beta_0 = -38.65 \\ \beta_1 = 60.77 \end{cases}$$

$$\rightarrow y = 60.77x - 38.65$$

سوال ۶-

الف)

طبقه اقتصادی:

$$H(X) = -\sum_{i=1}^{m} p_i log_2 p_i = -\left(\frac{1}{3} log_2\left(\frac{1}{3}\right) + \frac{1}{3} log_2\left(\frac{1}{3}\right) + \frac{1}{3} log_2\left(\frac{1}{3}\right)\right) \cong 1.58$$

حزب مورد علاقه:

$$H(Y) = -\sum_{i=1}^{m} p_i log_2 p_i = -\left(\frac{1}{2} log_2\left(\frac{1}{2}\right) + \frac{1}{2} log_2\left(\frac{1}{2}\right)\right) = 1$$

ب)

$$H(X,Y) = -\sum_i \sum_j P_{ij} log_2 P_{ij}$$
$$= -(0.1 * 3.321 + 0.175 * 2.514 + 0.225 * 2.152 + 0.25 * 2 + 0.1 * 3.321 + 0.15 * 2.736) = 2.49$$

$$I(X,Y) = H(X) + H(Y) - H(X,Y) = 1.58 + 1 - 2.49 = 0.09$$

ج) خیر، چون مقدار Mutual Information غیر از صفر است، بنابراین این دو متغیر از هم مستقل نمی‌باشند.

سوال ۷-

الف)

- انتخاب شاخصه: بعضی اوقات بهتر است که از همه شاخص‌ها استفاده نکنیم زیرا با شاخص‌های کمتر نیز می‌توان به نتیجه مطلوب رسید و نیازی به پیچیده‌تر کردن مدل نیست.
- تغییر حالت: در مواردی نیاز است که شکل داده‌ها تغییر کند تا امکان تحویل آن به مدل یادگیری وجود داشته باشد.
- انتخاب نمونه: آموزش را بر روی بخشی از داده‌ها انجام می‌دهیم که کیفیت بیشتری دارد.
- تمیز کردن داده: شامل حذف داده‌های ناقص، نویزها و داده‌های نامطلوب و غیرمرتبط می‌شود.
- نرمال‌سازی: چون ویژگی‌های هر داده مربوط به کمیت‌های مختلفی می‌شود، مقدار آن‌ها نیز ممکن است از لحاظ مقیاس تفاوت زیادی داشته باشد. بنابراین لازم است که مقیاس داده‌ها یکسان شود.

ب)

روش بیشینه‌کمینه:

$$min = 200, \qquad max = 1000$$

$$\left(\frac{200-min}{max-min}, \frac{300-min}{max-min}, \frac{400-min}{max-min}, \frac{600-min}{max-min}, \frac{1000-min}{max-min}\right) = (0, 0.125, 0.25, 1)$$

روش z-score:

$$mean = 500, \qquad std = 282.84$$

$$\left(\frac{200-mean}{std}, \frac{300-mean}{std}, \frac{400-mean}{std}, \frac{600-mean}{std}, \frac{1000-mean}{std}\right) = (-1.06, -0.7, -0.35, 0.35, 1.76)$$

سوال ۸-

الف)

$$f(\beta) = \left|\left|X\beta - y\right|\right|_2^2 + \alpha\left|\left|\beta\right|\right|_2^2 = (X\beta - y)^T(X\beta - y) + \alpha\beta^T\beta$$

$$\frac{\partial f(\beta)}{\beta} = 0$$

$$\rightarrow 2X^T(X\beta - y) + 2\alpha\beta = 0$$

$$\rightarrow (X^TX + \alpha I)\beta - X^Ty = 0$$

$$\rightarrow \beta = (X^TX + \alpha I)^{-1}X^Ty$$

ب)

$$Cost = (y - X\beta)^T(y - X\beta) + \lambda B^T B = \sum_{i=1}^{N}\left\{y_i - \sum_{j=0}^{M} B_j x_{ij}\right\}^2 + \lambda\sum_{j=0}^{M}\omega_j^2$$

$$\rightarrow \frac{\partial}{\partial\beta_j}(Cost) = -2\sum_{i=1}^{N} x_{ij}\left\{y_i - \sum_{k=0}^{M}\beta_k x_{ik}\right\} + 2\lambda B_j$$

$$\rightarrow \beta_j^{t+1} = \beta_j^t - \eta\left[-2\sum_{i=1}^{N} x_{ij}\left\{y_i - \sum_{k=0}^{M}\beta_k x_{ik}\right\} + 2\lambda B_j\right] = (1 - 2\lambda\eta)\beta_j^t + 2\eta\sum_{i=1}^{N} x_{ij}\left\{y_i - \sum_{k=0}^{M}\beta_k x_{ik}\right\}$$

ج) ترم منظم ساز اضافه شده را به صورت یک تابع تعریف می‌کنیم و سپس اثبات می‌کنیم که اگر $\alpha \geq 0$ آنگاه این تابع محدب است:

$$f(\beta) = \alpha||\beta||_2^2$$

$$\alpha \geq 0 \rightarrow f(\theta\beta_1 + (1-\theta)\beta_2) = \alpha||\theta\beta_1 + (1-\theta)\beta_2||_2^2 \leq \left(\theta\alpha||\beta_1||_2^2 + (1-\theta)\alpha||\beta_2||_2^2\right)$$
$$= \theta f(\beta_1) + (1-\theta)f(\beta_2)$$

پیاده‌سازی

قسمت اول:

۱-

	id	sex	birth_year	country	region	infection_reason	infected_by	confirmed_date	state
0	1	female	1984.0	China	filtered at airport	visit to Wuhan	NaN	1/20/2020	released
1	2	male	1964.0	Korea	filtered at airport	visit to Wuhan	NaN	1/24/2020	released
2	3	male	1966.0	Korea	capital area	visit to Wuhan	NaN	1/26/2020	released
3	4	male	1964.0	Korea	capital area	visit to Wuhan	NaN	1/27/2020	released
4	5	male	1987.0	Korea	capital area	visit to Wuhan	NaN	1/30/2020	released
...	...	...	...	...	...	...	...	...	...
171	172	female	1997.0	Korea	Gyeongsangbuk-do	NaN	NaN	2/24/2020	isolated
172	173	male	1949.0	Korea	Daegu	NaN	NaN	2/24/2020	deceased
173	174	female	1958.0	Korea	Gyeongsangbuk-do	NaN	NaN	2/24/2020	isolated
174	175	male	1997.0	Korea	Gyeongsangbuk-do	NaN	NaN	2/24/2020	isolated
175	176	female	1950.0	Korea	capital area	NaN	NaN	2/24/2020	isolated

176 rows × 9 columns

۲- هر سطر جدول مربوط به یک بیمار کووید-۱۹ بوده و هر ستون مشخصات و ویژگی‌های آن بیمار را نمایش می‌دهد. تعداد داده‌ها ۱۷۶تا بوده و نام ستون‌ها عبارتند از:

id, sex, birth_year, country, region, infection_reason, infected_by, confirmed_date, state

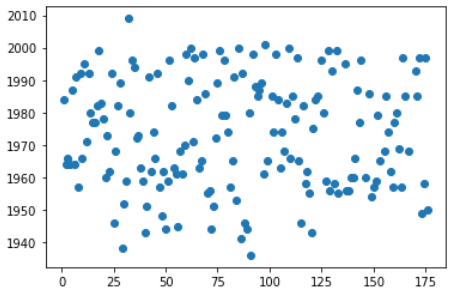
۳-

```
mean: 1973.3855421686746
max: 2009.0
std: 16.981443682011555
```

۴- بله، مقادیر null در داده‌ها وجود دارد. با استفاده از متد dropna() می‌توان داده‌هایی که برخی از ویژگی‌های آن‌ها null است را حذف کرد.

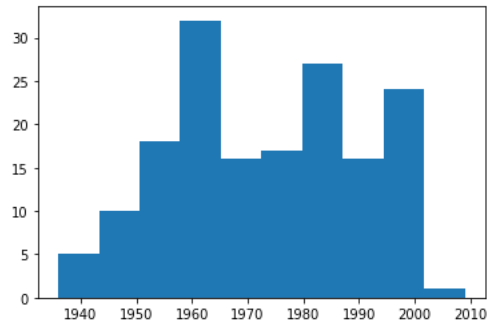
۵-

Scatter plot:



Matrix plot:

Histogram plot:



۶- بله، یکی از موارد ابتلا مربوط به فردی است که متولد سال ۲۰۰۹ می‌باشد. با توجه به اختلاف سن این فرد با بقیه موارد می‌توان گفت که این داده یک outlier است. برای حل این مشکل می‌توان بررسی کرد که آیا این داده صحت دارد یا خیر و در صورت نادرستی باید آن را از مجموعه داده‌ها حذف کرد.

قسمت دوم:

رگرسیون خطی انجام شد. میزان خطا به طور میانگین (Mean Absolute Error) برابر با ۳.۲۲ نمره بود.