به نام خدا

دانشگاه تهران

دانشکدگان فنی

دانشکده مهندسی برق و کامپیوتر

درس داده کاوی

تمرین اول

نام و نام خانوادگی : حسین سیفی

شماره دانشجویی : ۸۱۰۱۰۰۳۸۶

اسفند ۱۴۰۰

فهرست

۱. سوال ۱ ..... ۳

۱-۱. الف ۳

۱-۲. ب ۳

۲. سوال ۲ ..... ۳

۲-۱. الف ۳

۲-۲. ب ٤

۲-۳. ج ٥

۲-۴. د ٦

۲-۵. ه ٦

۳. سوال ۳ ..... ۷

۳-۱. الف ۷

۳-۲. ب ۸

۳-۳. ج ۸

۱. سوال ۱

۱-۱. الف

Attribute	Type	Discrete or Continuous
سرعت خودرو بر اساس MPH	Ratio-Scaled	Continuous
شدت بارش با مقادیربدون بارش ، بارش متناوب و بارش بی وقفه	Ordinal	Discrete
تجزیه نور به رنگ‌های رنگین کمان	Nominal	Discrete
اندازه گیری هوش افراد با IQ	Ratio-Scaled	Discrete
عدد بارکد اجناس در سوپرمارکت	Nominal	Discrete

۱-۲. ب

	Nominal	Ordinal	Interval-Scaled	Ratio-Scaled
Histogram			*	*
Pie chart	*	*		
Box-Plot			*	*
Bar chart	*	*		

۲. سوال ۲

۲-۱. الف

میانگین خصیصه A و B به شکل زیر محاسبه می‌شود:

$$avg(A) = \frac{\sum_{i=0}^{19} Ai}{count(A)} = \frac{2839.4}{20} = 141.97$$

$$avg(B) = \frac{\sum_{i=0}^{19} Bi}{count(B)} = \frac{243.73}{20} = 12.1865$$

برای به دست آوردن چارک اول و سوم می‌توانیم مقادیر هر خصیصه را مرتب کنیم:

A : 0.04, 0.31, 0.78, 6.21, 11.7, 42.5, 43.02, 58.1, 72, 75.3, 95, 110.02, 111, 111.87, 121.2, 145, 265.6, 311.54, 598.23, 659.98

B : 0.53, 1.5, 2.3, 3.9, 4.2, 5.6, 6.7, 7.43, 7.8, 8.6, 11.3, 12.1, 12.9, 13.2, 15.8, 17.2, 19.8, 23.9, 34.17, 34.8

با توجه به اینکه تعداد مقادیر هر دو خصیصه زوج و برابر ۲۰ می‌باشد ، چارک اول برابر میانگین مقادیر پنجم و ششم (بین مقادیر مرتب شده) و چارک سوم برابر میانگین مقادیر پانزدهم و شانزدهم می‌باشد:

$$FirstQuartile(A) = \frac{11.7 + 42.5}{2} = 27.1 \text{ و } ThirdQuartile(A) = \frac{121.5 + 145}{2} = 133.25$$

$$FirstQuartile(B) = \frac{4.2 + 5.6}{2} = 4.9 \text{ و } ThirdQuartile(B) = \frac{15.8 + 17.2}{2} = 16.5$$

با استفاده از مقادیر خصیصه‌ها و مقدار میانگین هر خصیصه ، مقدار انحراف معیار را با فرمول زیر محاسبه می‌کنیم:

$$\sigma = \sqrt{\frac{\sum_{i=0}^{N-1}(Xi - \mu)^2}{N}}$$

$$\sigma(A) = \sqrt{\frac{(0.04 - 141.97)^2 + \cdots + (659.98 - 141.97)^2}{20}} = \sqrt{\frac{657686.2148}{20}} = \sqrt{32884.31} = 181.34$$

$$\sigma(B) = \sqrt{\frac{(0.53 - 12.1865)^2 + \cdots + (34.8 - 12.1865)^2}{20}} = \sqrt{\frac{1838.82}{20}} = \sqrt{91.941} = 9.5885$$

۲-۲. ب

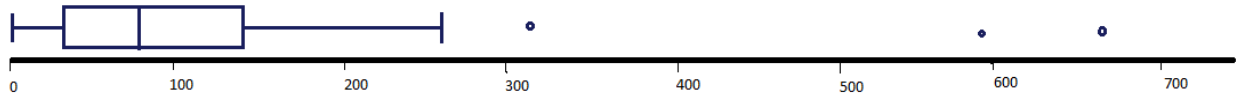
برای رسم نمودار جعبه‌ای به ۵ مقدارمیانه ، چارک اول ، چارک سوم ، حداقل مقدار و حداکثر مقدار هر خصیصه نیازمندیم.این مقادیر برای خصیصه A به شرح زیر می‌باشند:

Median(A) = (75.3+95)/2 = 85.15
Minimum(A)= 0.04
Maximum(A)= 659.98
Q3(A)= 133.25
Q1(A)= 27.1

دیگر مقادیر مورد نیاز برای رسم نمودار به کمک مقادیر فوق قابل محاسبه است:

IQR(A) = Q3(A) – Q1(A) = 133.25 – 27.1 = 106.15
Low Outlier Treshold(B) = 27.1 – 1.5 * 106.15 = -132.65
High Outlier Treshold(B) = 133.25 + 1.5 * 106.15 = 292.475

با توجه به اینکه داده‌های جمع آوری شده برای خصیصه A همگی مثبت هستند و همچنین مقداری کمتری از آستانه پایین outlier وجود ندارد پس حد پایین نمودار همان حداقل مقدار خصیصه می‌باشد. از طرفی دیگر این خصیصه مقادیری بزرگتر آستانه بالا outlier وجود دارد و در نتیجه این outlierها به شکل نقاطی پراکنده نمایش داده می‌شوند. در نهایت ، نمودار جعبه‌ای خصیصه A به طور تقریبی به شکل زیر درمی‌آید:



مقادیر میانه ، چارک اول ، چارک سوم ، حداقل مقدار و حداکثر مقدار برای خصیصه B به شرح زیر می‌باشند:

Median(B) = (8.6+11.3)/2 = 9.95
Minimum(B)= 0.53
Maximum(B)= 34.8
Q3(B)= 16.5
Q1(B)= 4.9

دیگر مقادیر مورد نیاز برای رسم نمودار به کمک مقادیر فوق قابل محاسبه است:

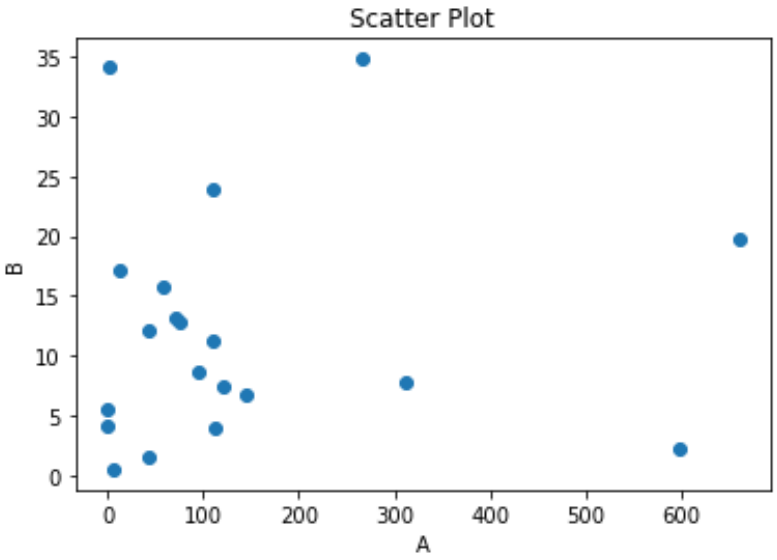
IQR(B) = Q3(B) – Q1(B) = 16.5 – 4.9 = 11.6
Low Outlier Treshold(B) = 4.9 – 1.5 * 11.6 = -12.5
High Outlier Treshold(B) = 16.5 + 1.5 * 11.6 = 33.9

با توجه به داده‌های فوق و توضیحات ذکر شده برای نمودار جعبه‌ای خصیصه A، نمودار جعبه‌ای خصیصه B به شکل زیر درمی‌آید:

با توجه به اینکه نمودار جعبه‌ای خصیصه A بازه IQR بزرگتری نسبت به خصیصه B دارد و خصیصه A طول Whiskerهای بیشتری نسبت به B دارد و همچنین بیشتر بودن چشمگیر انحراف معیار A نسبت به B نتیجه می‌گیریم که مقادیر داده‌ها در خصیصه A نسبت به خصیصه B پراکنده‌تر هستند.

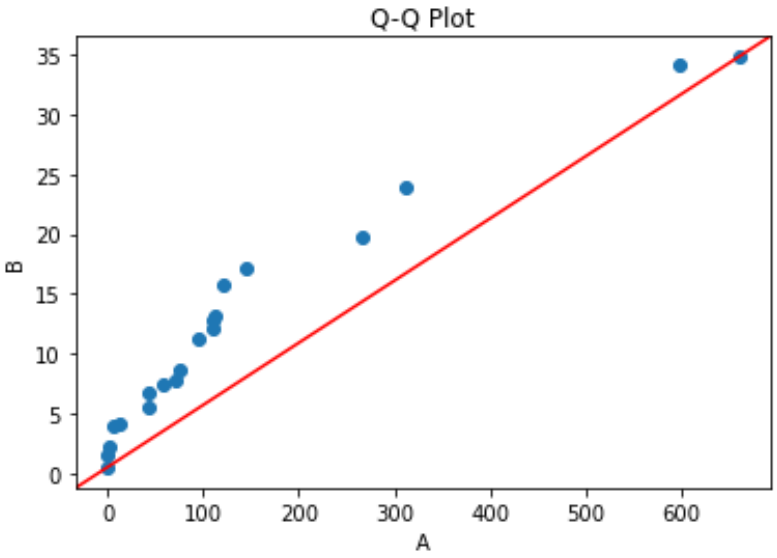
۳-۲. ج

برای رسم نمودار پراکندگی از زوج (A_i,B_i) با همین ترتیب ذکر شده در سوال استفاده می‌کنیم. نتیجه به شکل زیر درمی‌آید:



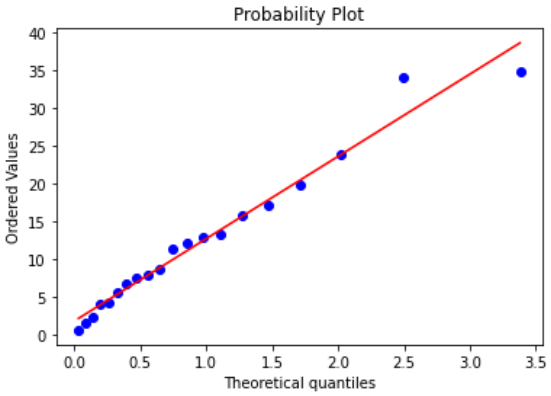
۲-۴. د

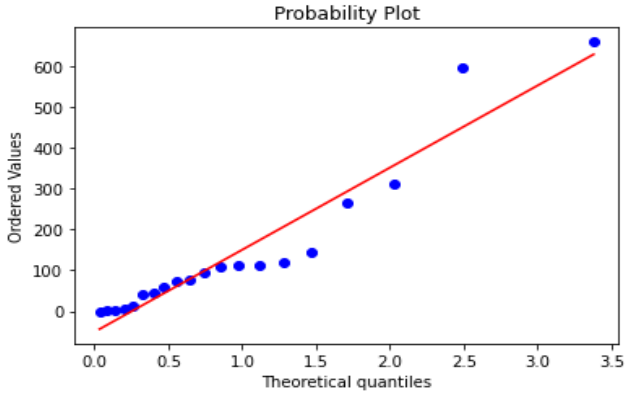
برای رسم نمودار Q-Q plot خصیصه A نسبت به خصیصه B از زبان پایتون استفاده کرده‌ایم. این پیاده سازی بدین صورت انجام گرفته است که با توجه به اینکه هر دو خصیصه ۲۰ مقدار دارند می‌توان لیست مقادیر هر خصیصه را مرتب کرد و مقدار هر عضو از خصیصه A همراه با مقدار متناظر با اندیس آن در خصیصه B مختصات یک نقطه را نشان می‌دهند.(عضو iام (0<i<21) در لیست مرتب شده هر خصیصه نشان دهنده 5i-Quantile است.) همچنین خطی مورب برای نمایش وابستگی داده‌های محور عمودی و افقی رسم شده است. نمودار رسم شده به شکل زیر می‌باشد:



۲-۵. ه

نمودارهای زیر مقایسه داده‌های نمونه A و B را با داده‌های تئوری از توزیع نمایی نشان می‌دهند.(نمودار سمت راست مربوط به خصیصه A و نمودار چپ مربوط به خصیصه B می‌باشد) همانطور که مشخص است ، مقادیر خصیصه B توزیع نمایی دارند اما خصیصه A از این توزیع پیروی نمی‌کند.(توزیع‌های دیگری نیز بررسی شدند اما مقادیر این خصیصه از هیچ توزیع شناخته شده‌ی دیگری پیروی نمی‌کند) این موضوع به همراه بررسی نمودارهای رسم شده در بخش ب ، ج و د همین سوال نشان می‌دهد که این دو خصیصه هم‌بستگی ندارند.





۳. سوال ۳

۳-۱. الف

با استفاده از زبان برنامه نویسی پایتون توابعی برای محاسبه فواصل منهتن ، اقلیدسی ، سوپریمم و شباهت کسینوسی تعریف شده‌اند تا این مقادیر را برای زوج کلمه پرسش و هر کدام از کلمات X_1 تا X_5 محاسبه کند. نتیجه محاسبات و رتبه بندی فاصله‌ها به شرح زیر می‌باشد:

Rank	**Euclidean**	**Manhattan**	**Supremum**	**Cosine Similarity**
1	X_1	X_1	X_1	X_1
2	X_4	X_4	X_3 , X_4	X_3
3	X_3	X_3	X_2 , X_5	X_4
4	X_5	X_5	-	X_2
5	X_2	X_2	-	X_5

مقادیر این فواصل نیز در جدول زیر به تفکیک قابل مشاهده است:

Rank	**Euclidean**	**Manhattan**	**Supremum**	**Cosine Similarity**
1	0.141	0.2	0.1	0.999991
2	0.224	0.3	0.2	0.999969
3	0.283	0.4	0.6	0.999028
4	0.608	0.7	-	0.995752
5	0.671	0.9	-	0.965363

فرمول فواصل استفاده شده در کد پایتون به شکل زیر می‌باشد:

- فاصله اقلیدسی:

$$Euclidean\ Distance(x,y) = \sqrt{\left(A_{1,x} - A_{1,y}\right)^2 + (A_{2,x} - A_{2,y})^2}$$

- فاصله منهتن:

$$Manhattan\ Distance(x,y) = \left|A_{1,x} - A_{1,y}\right| + |A_{2,x} - A_{2,y}|$$

- فاصله سوپریمم:

$$Supremum\ Distance(x,y) = \max\left(\left|A_{1,x} - A_{1,y}\right| \ , \ \left|A_{2,x} - A_{2,y}\right|\right)$$

- شباهت کسینوسی:

$$Cosine\ Similarity(x,y) = \frac{\left(A_{1,x} * A_{1,y}\right) + (A_{2,x} * A_{2,y})}{\sqrt{{A_{1,x}}^2 + {A_{2,x}}^2} * \sqrt{{A_{1,y}}^2 + {A_{2,y}}^2}}$$

۲-۳. ب

برای نرمال سازی داده‌های X_1 تا X_5 نیاز به مقادیر میانگین و انحراف معیار برای هر یک از خصیصه‌های A_1 و A_2 داریم که با استفاده از فرمول‌های ذکر شده در بخش الف سوال ۲ قابل محاسبه است. مقدار میانگین و انحراف معیار برای خصیصه A_1 به ترتیب برابر ۱.۵۶ و ۰.۲۵۷۶ و میانگین و انحراف معیار برای خصیصه A_2 برابر ۱.۵۸ و ۰.۳۱۸۷ می‌باشند. سپس هر کدام از مقادیر خصیصه‌ها به کمک فرمول زیر نرمال سازی می‌شود:

$$Z - Score(x) = \frac{x - \mu}{\sigma}$$

مجموعه کلمات X_1 تا X_5 پس از نرمالسازی به صورت بردارهای دو بعدی زیر درمی‌آید:

	A_1	**A_2**
X_1	-0.2328	0.3764
X_2	1.7075	1.0039
X_3	0.1552	0.6902
X_4	-1.3970	-0.2509
X_5	-0.2328	-1.8196

اگر بردار پرس و جو را نرمال نکنیم فاصله اقلیدسی که بین این بردار و بردار کلمات به دست می‌آید مقدار دقیقی را نشان نمی‌دهد و در بعضی موارد ، لیست مرتب شده فواصل ترتیب درستی ندارند و قابل اتکا نیستند. به همین دلیل برای محاسبه هر نوع فاصله باید بردار پرس و جو نیز به کمک فرمول Z-score نرمال شود . فرم نرمال شده پرس و جو به شکل X=(-0.620,0.062) می‌باشد.

این بخش سوال نیز به کمک زبان برنامه نویسی پایتون از صفر پیاده سازی شده است و ترتیب فاصله اقلیدسی کلمات X_1 تا X_5 با کلمه پرس و جو X و مقادیر فواصل به شرح زیر می‌باشند:

Rank	**Word Number**	**Euclidean Distance**
1	X_1	0.499
2	X_4	0.837
3	X_3	0.998
4	X_5	1.922
5	X_2	2.511

۳-۳. ج

- برای محاسبه عدم شباهت بین متغیرهای ratio-scaled می‌توان بر روی مقادیر این متغیرها تابع لگاریتم را اعمال کرد و سپس با اعداد به دست آمده که مقداری حقیقی دارند مانند متغیرهای Interval-scaled برخورد کرد و از فواصل اقلیدسی ، منهتن ، سوپریمم و شباهت کسینوسی استفاده کرد.
- روش دیگر این است که بدون تغییر داده‌ها ، با آنان مانند داده‌های Interval-Scaled رفتار کنیم که اغلب راه حل مناسبی نیست و به درستی پاسخ نمی‌دهد.