



# به نام خدا

# پاسخنامه تمرین اول

**استاد درس:**

**دکتر مریم امیر مزلقانی**

نیم سال اول ۱۴۰۲-۱۴۰۳

راه ارتباطی:

Aut.DataMining.Fall@gmail.com

**سوال ۱.**

( پاسخ بسته به تعریف شما از مسئله می تواند متفاوت باشد)

۱. تعریف مسئله: در این مرحله، باید مسئله‌ای که قرار است با داده کاوی حل شود، تعریف شود. به عنوان مثال، ممکن است میخواهید برای شرکت تولید دارو، یک مدل پیش‌بینی تقاضا برای داروها بسازید تا بتوانید تولید و توزیع بهینه را برنامه‌ریزی کنید.
۲. جمع‌آوری داده: در این مرحله، باید داده‌های لازم جهت حل مسئله جمع‌آوری شوند. ممکن است شما نیاز داشته باشید به داده‌های مربوط به فروش قبلی داروها، اطلاعات درباره مشتریان و یا هر داده دیگری که ممکن است برای پیش‌بینی تقاضا مفید باشد.
۳. پیش‌پردازش داده: در این مرحله، داده‌های جمع‌آوری شده نیاز به پیش‌پردازش دارند تا قابل استفاده در مدل‌های داده کاوی باشند. این مرحله شامل تمیزکاری داده‌ها، حذف داده‌های نامناسب یا اشتباه، پر کردن مقادیر خالی و استخراج ویژگی‌های مهم است.
۴. انتخاب و آماده‌سازی مدل: در این مرحله، باید یک مدل داده کاوی را انتخاب کنید که بتواند مسئله را حل کند. ممکن است از روش‌هایی مانند رگرسیون، شبکه‌های عصبی، درخت تصمیم و یا الگوریتم‌های دیگر استفاده کنید .سپس مدل را آموزش داده و بهبود دهید تا بهترین عملکرد را در پیش‌بینی تقاضا داشته باشد.
۵. ارزیابی مدل: در این مرحله، عملکرد مدل را با استفاده از معیارهای مناسبی مانند دقت، صحت، پیشبینی‌های صحیح و خطاهای مدل ارزیابی کنید. این مرحله به شما کمک می‌کند تا ببینید که آیا مدل به طور قابل قبولی تقاضاها را پیش‌بینی می‌کند یا نه.
۶. استفاده از مدل: پس از آموزش و ارزیابی مدل، می‌توانید آن را برای پیش‌بینی تقاضاهای آینده استفاده کنید. از این پیش‌بینی‌ها می‌توانید در برنامه‌ریزی تولید و توزیع داروها استفاده کنید تا موجودی داروها را بهینه کنید و به مشتریان خدمات بهتری ارائه دهید.

از طرف دیگر، در هر مرحله می‌توانید فرآیند را بهبود دهید و مدل‌ها و الگوریتم‌های جدید را امتحان کنید تا عملکرد بهتری بدست آورید. همچنین می‌توانید داده‌های جدید را به مرور زمان جمع‌آوری کنید و مدل را به‌روزرسانی کنید تا به دقت بیشتری در پیش‌بینی تقاضا برسید

**سوال ۲.**

- Noise (نویز): نویز به مقادیر تصادفی و بدون ساختار در داده‌ها اشاره دارد. نویز معمولاً به صورت تصادفی و بدون روند خاصی در داده‌ها حضور دارد و ممکن است اطلاعات غیرضروری و ناخواسته را به داده‌ها اضافه کند. نویز معمولاً نتیجه‌ی عوامل مختلفی مانند خطاهای اندازه‌گیری، اشکال در فرآیند جمع‌آوری داده و تداخلات اندازه‌گیری است. معایب نویز شامل کاهش دقت تحلیل‌ها، تأثیر منفی بر روی مدل‌های یادگیری ماشین و مشوق کندی فرآیند تحلیل داده می‌شود.

- Outlier (نقطه‌ی نامعمول): نقطه‌ی نامعمول یا outlier به مقداری در داده‌ها اشاره دارد که به طور قابل مشاهده و برجسته از سایر داده‌ها متمایز است و از الگوها و روندهای معمول داده‌ها خارج می‌شود. برخلاف نویز که به صورت تصادفی در داده‌ها وجود دارد، outlier معمولاً نتیجه‌ی خطاها، نقص‌ها یا رویدادهای استثنایی در داده‌ها است. معایب outlier شامل تحریف تحلیل‌ها، کاهش دقت مدل‌ها و الگوریتم‌های یادگیری ماشین و نگرانی‌های احتمالی در مورد صحت و قابل اعتماد بودن داده است.

۱) پیدا کردن outlier به عنوان یک مسئله مهم در داده کاوی، بسیاری از سودها و فواید را به همراه دارد. در زیر تعدادی از این سودها را بررسی می‌کنیم:

- شناسایی خطاها و نقص‌ها: outlier ها ممکن است نشان دهنده‌ی خطاها و نقص‌ها در فرآیند جمع‌آوری داده باشند. با تحلیل و شناسایی این نقاط، می‌توان مشکلات را تشخیص داده و بهبود فرآیند جمع‌آوری داده را انجام داد.

- تحلیل صحیح داده‌ها: وجود outlier در داده‌ها می‌تواند تحلیل‌های غلط و نادرست را تولید کند. با حذف یا تصحیح این نقاط نامعمول، می‌توان مدل‌ها و الگوریتم‌ها را بهبود بخشید و تحلیل دقیق‌تری روی داده‌ها انجام داد.

- شناسایی الگوهای جدید: در برخی موارد، outlier ها می‌توانند نشان‌دهنده‌ی الگوها و روندهای جدید در داده‌ها باشند. شناسایی این الگوها می‌تواند به کشف دانش جدید و اطلاعات مفیدی در مورد داده‌ها کمک کند.

- بهبود دقت مدل‌ها: وجود outlier در داده‌ها می‌تواند تأثیر بسیار زیادی بر روی دقت مدل‌ها و الگوریتم‌های یادگیری ماشین داشته باشد. با حذف این نقاط نامعمول، می‌توان دقت مدل‌ها را بهبود بخشید و عملکرد بهتری را در پیش‌بینی و تحلیل داده‌ها ارائه داد.

پیدا کردن outlier در داده کاوی به صورت گسترده در بسیاری از حوزه‌ها و مسئله‌های مختلف از جمله Fraud Detection استفاده می‌شود. مثالی از مسئله Fraud Detectionمی‌تواند تشخیص تقلب در تراکنش‌های اینترنتی باشد. در این مسئله، هدف اصلی پیدا کردن outlier ها و تشخیص تراکنش‌های نامعمول و تقلبی است.

به طور مثال، فرض کنید بیشتر تراکنش‌ها در یک منطقه خاص انجام می‌شود و مبلغ متوسط تراکنش‌ها در آن منطقه به طور معمول در حدود ۱۰۰ دلار است. اگر یک تراکنش با مبلغ ۱۰۰۰ دلار ثبت شود، این تراکنش به عنوان یک outlier شناخته می‌شود. با پیدا کردن این نقطه نامعمول، می‌توان تراکنش‌های تقلبی را تشخیص داد و اقدامات لازم را برای جلوگیری از تقلب انجام داد.

۲) Noise و outlier هر دو به عنوان اشکال و نقص‌های ممکن در داده‌ها شناخته می‌شوند، اما با خصوصیات متفاوتی همراه هستند.

معایب نویز:

۱. تشویش داده: نویزها ممکن است داده‌ها را تغییر دهند و اطلاعات مفید را از بین ببرند، این موضوع می‌تواند باعث کاهش دقت و قابلیت اطمینان در تحلیل داده‌ها شود.

۲. پیچیدگی تحلیل: وجود نویزها می‌تواند تحلیل داده‌ها را پیچیده کند و نیاز به روش‌های پیشرفته‌تری برای استخراج اطلاعات مفید از داده‌ها ایجاد کند.

معایب اوتلایر:

۱. تأثیر منفی بر آماره‌ها: وجود اوتلایرها می‌تواند تأثیر زیادی بر معیارهای آماری مانند میانگین و واریانس داشته باشد و تفسیر نادرستی از داده‌ها ارائه دهد.

۲. اشتباهات در مدل‌سازی: اگر اوتلایرها در مدل‌های آماری و یا ماشینی در نظر گرفته نشوند، ممکن است مدل‌ها نتایج نادرستی تولید کنند و در نتیجه پیش‌بینی‌ها و تحلیل‌ها ناقص شوند.

۳. کاهش دقت: اوتلایرها ممکن است به عنوان داده‌های معمول در نظر گرفته شوند و تحلیل‌ها و پیش‌بینی‌ها را تحت تأثیر قرار دهند، که منجر به کاهش دقت و قابلیت اعتماد در مدل‌ها و سیستم‌های تحلیلی می‌شود.

نکته:

روش Z-score یکی از روش‌های متداول در آمار و احتمالات است که برای محاسبه و ارزیابی فاصله یک داده نسبت به میانگین مورد انتظار و واحد انحراف استاندارد استفاده می‌شود. با استفاده از این روش، می‌توانید ببینید که یک داده چقدر از میانگین فاصله دارد و آیا این فاصله نسبت به توزیع داده‌ها عادی است یا نه.

فرمول محاسبه Z-score برای یک داده به شکل زیر است:

$$Z = (X - \mu) / \sigma$$

که:

- Zنشان‌دهنده Z-score است.
- Xمقدار داده است که قصد داریم Z-score آن را محاسبه کنیم.
- $\mu$میانگین مورد انتظار داده‌ها است.
- $\sigma$واحد انحراف استاندارد داده‌ها است.

مقدار Z-score نشان می‌دهد که یک داده چند واحد استاندارد از میانگین فاصله دارد. به طور معمول، اگر مقدار Z-score بیشتر از ۳ یا کمتر از ۳- باشد، معمولاً به عنوان یک اوتلایر در نظر گرفته می‌شود.

با استفاده از روشZ-score ، می‌توانید نویزها و اوتلایرها را در داده‌ها تشخیص دهید. اگر یک داده مقدار Z-score بالایی داشته باشد، احتمالاً یک اوتلایر است. به عبارت دیگر، اگر فاصله یک داده از میانگین به طور غیرمعمول بزرگ باشد، مقدار Z-score بیشتر از مقادیر آستانه (مثلاً ۳) خواهد بود و ما می‌توانیم آن را به عنوان یک اوتلایر شناسایی کنیم.

به طور مشابه، اگر Z-score یک داده به صورت معمول در بازه (۳,۳-) باشد، می‌توانیم آن را به عنوان یک داده عادی و بدون نویز در نظر بگیریم.

**سوال ۳.**

الف)

داده نسبتی:(Ratio Data)

- مثال ۱: وزن افراد به کیلوگرم. به عنوان مثال، وزن یک فرد می‌تواند ۷۰ کیلوگرم باشد. دلیل: داده نسبتی برخلاف سایر انواع داده‌ها، دارای مبنای مطلق است. می‌توان از عملیات ریاضی مثل جمع، تفریق، ضرب و تقسیم برای این نوع داده‌ها استفاده کرد. به این معنی که می‌توان نسبت‌های وزن بین افراد را محاسبه کرد. مثلاً، وزن یک فرد دو برابر وزن دیگری باشد.
- مثال ۲: تعداد قطره‌های بارش در یک روز. به عنوان مثال، تعداد قطره‌های بارش در یک روز ممکن است ۱۰۰ قطره باشد. دلیل: داده نسبتی امکان محاسبه نسبت مقادیر واقعی را فراهم می‌کند. می‌توان نسبت تعداد قطره‌ها در یک روز به دو روز مختلف را محاسبه کرد و بگوییم که در روز اول تعداد قطره‌ها دو برابر روز دوم بوده است.

داده فاصله‌ای:(Interval Data)

- مثال ۱: دمای هوا بر حسب سانتیگراد. به عنوان مثال، دمای هوا ممکن است ۲۵ درجه سانتیگراد باشد. دلیل: داده فاصله‌ای دارای مبنای نسبتی نیست، اما می‌توان با استفاده از عملیات ریاضی تفاوت بین دو مقدار را محاسبه کرد .مثلاً، تفاوت دمای ۲۵ درجه و ۱۵ درجه برابر با ۱۰ درجه است.
- مثال ۲: زمان در ساعت و دقیقه. به عنوان مثال، زمان ممکن است ۱۴:۳۰ باشد. دلیل: در داده فاصله‌ای، می‌توان تفاوت زمان بین دو رویداد را محاسبه کرد. مثلاً، می‌توان گفت که زمان بین دو رویداد ۱۴:۳۰ و ۱۵:۰۰ برابر با ۳۰ دقیقه است.

داده ترتیبی:(OrdinalData)

- مثال ۱: رتبه‌بندی دانشجویان بر اساس نمره. به عنوان مثال، دانشجویان می‌توانند در رتبه اول، دوم و سوم قرار گیرند. دلیل: داده ترتیبی حاوی اطلاعات درباره ترتیب و رتبه‌بندی دارد. می‌توان از عملیات مقایسه برای مقایسه این داده‌ها استفاده کرد. مثلاً، می‌توان گفت که دانشجویی که در رتبه اول است، از نظر نمره بالاتری نسبت به دانشجویان دیگر دارد.
- مثال ۲: میزان رضایتمندی مشتریان بر اساس مقیاس ۱ تا ۵. به عنوان مثال، مشتریان می‌توانند رضایتمندی خود را با اعداد ۱ تا ۵ اعلام کنند. دلیل :داده ترتیبی دارای ترتیب است و می‌توان مقایسه‌ای بین داده‌ها انجام داد .می‌توان نتیجه گرفت که رضایتمندی مشتریانی که اعلام کرده‌اند ۵ برابر با رضایتمندی مشتریانی است که اعلام کرده‌اند ۳.

داده نامی:(Nominal Data)

- مثال ۱: رنگ ماشین‌ها. به عنوان مثال، رنگ ماشین‌ها می‌تواند سفید، مشکی و قرمز باشد. دلیل: داده نامی نمی‌تواند مقادیر را مقایسه کند یا به ترتیب بندی برساند. آنها به صورت دسته‌بندی شده هستند و می‌توان فقط اطلاعات وجود یا عدم وجود هر دسته را مشخص کرد.
- مثال ۲: جنسیت افراد. به عنوان مثال، جنسیت ممکن است مرد یا زن باشد .دلیل: داده نامی برای دسته‌بندی و شناسایی موارد استفاده می‌شود، اما نمی‌توان مقداری را به صورت ریاضی مورد استفاده قرار داد. به عنوان مثال، نمی‌توان گفت که مقدار زن برابر با دو برابر مقدار مرد است

ب)

شماره دانشجویی:
مسئله: بررسی عملکرد تحصیلی دانشجویان در یک دانشگاه.
توضیح: شماره دانشجویی به عنوان یک شناسه یکتا برای هر دانشجو استفاده می‌شود. این ویژگی می‌تواند در مطالعات مربوط به عملکرد تحصیلی دانشجویان بسیار مفید و تاثیرگذار باشد. با استفاده از شماره دانشجویی، می‌توان داده‌های مربوط به هر دانشجو را به صورت منحصر به فرد ردیابی کرد، مقایسه‌های زمانی انجام داد و تغییرات در عملکرد تحصیلی هر دانشجو را بررسی کرد.

جنسیت:
مسئله: تحلیل عوامل موثر بر نتایج پژوهش در حوزه بهداشت و پزشکی.
توضیح: جنسیت به عنوان یک ویژگی بیولوژیکی و اجتماعی می‌تواند در تحلیل عوامل موثر بر نتایج پژوهش در حوزه بهداشت و پزشکی تاثیرگذار باشد .تفاوت‌های بین جنسیت‌ها می‌تواند درک بهتری از عوامل خطر و عوارض بیماری‌ها، تفاوت در عوامل ایمنی و پاسخ به درمان، و تأثیرات داروها بر جنسیت‌ها را فراهم کند. این ویژگی می‌تواند به محققان در تصمیم‌گیری‌های مربوط به طراحی پژوهش، تجزیه و تحلیل داده‌ها و ارائه نتایج کمک کند .همچنین، ارائه گزارش‌های جداگانه برای هر جنسیت در مطالعات بهداشتی و پزشکی نقش مهمی در بهبود تفهیم و درک عواقب سلامتی دارد

ج)

اسمی :برای داده های اسمی مد قابل تعریف است. چرا که تعداد مقادیر مختلف قابل شمارش است. میانه نیز قابل تعریف نیست زیرا ترتیب برای این داده مشخص نیست و طبق تعریف میانه نیاز به یک ترتیب مشخص برای مرتب سازی داده ها و پس یافتن میانه آن ها داریم. میانگین نیز قابل تعریف نیست چرا که این مقادیر گسسته هستند و میانگین بر ای آنها بی معناست. به عبارت دیگر برای یافتن میانگین نیاز به جمع و تقسیم داریم که برای داده های اسمی تعریف نمی شود.

ترتیبی :برای داده های ترتیبی نیز مد قابل تعریف است چرا که به وضوح می توان تعداد مقادیر مختلف را محاسبه کرد .برخالف داده های اسمی با توجه به اینکه در داده های ترتیبی برای داده ها تر تیب مشخصی داریم و میتوانیم آنها را مرتب کنیم لذا میانه

برای آنها تعریف می شود. در مورد میانگین نیز مانند داده های اسمی با توجه به اینکه برای این نوع داده جمع و تفریق و تقسیم تعریف نمی شود امکان محاسبه میانگین وجود ندارد.

بازه ای :برای مقادیر بازه ای مد قابل تعریف است چرا که مقادیر مختلف قابل شمارش هستند. همچنین با توجه به اینکه این نوع داده ها ترتیب دارند میانه برای آنها قابل محاسبه است .میانگین برای داده های بازه ای مانند تاریخ را نمی توان مستقیماً محاسبه کرد زیرا یک مقدار عددی نیست ، با این حال می توان تاریخ را به مقدار عددی مانند تعداد روزهای پس از یک تاریخ مرجع خاص تبدیل کرد و سپس میانگین آن مقادیر عددی را محاسبه کرد .

نرخی :با توجه به اینکه این مقادیر قابل شمارش هستند در نتیجه مد برای آنها مانند سایر انواع داده قابل تعریف است .در مورد میانگین نیز با توجه به تعریف شدن تربیت برای ای ن نوع داده میتوان میانه را برای آنها تعریف کرد .همچنین با توجه به تعریف شدن ضرب و تقسیم و جمع و تفریق برای آنها میتوان به وضوح با جمع کردن داده ها و تقسیم کردن آنها به تعداد میانگین آن ها را تعریف و محاسبه کرد.

**سوال ۴.**

الف)

۱. One-Hot Encoding:

یک روش رایج برای تبدیل داده‌ها به بردارهای باینری است. این روش برای متغیرهای دسته‌ای با مقادیر گسسته استفاده می‌شود. در این روش، برای هر مقدار ممکن در متغیر دسته‌ای، یک بردار باینری به طول تعداد مقادیر ممکن در نظر گرفته می‌شود. در این بردار، تمام عناصر به جز عنصر متناظر با مقدار واقعی، صفر هستند و عنصر متناظر با مقدار واقعی برابر یک است

۲. Label Encoding:

در روش Label Encoding، هر داده به یک عدد صحیح تبدیل می‌شود. برای این کار، تمام مقادیر ممکن برای یک ویژگی مشخص شناسایی می‌شوند و به هر یک از آن‌ها یک عدد منحصر به فرد نسبت داده می‌شود. این روش برای دسته‌بندی‌های دو دسته‌ای و دسته‌بندی‌های چند دسته‌ای با ترتیب مناسب است.

ب)

| ID | Color | Shape |
|---|---|---|
| 1 | 2 | 1 |
| 2 | 1 | 1 |
| 3 | 2 | 2 |
| 4 | 3 | 3 |

| ID | Red | Blue | Green | Tri | Rou | Squ |
|---|---|---|---|---|---|---|
| 1 | 0 | 1 | 0 | 1 | 0 | 0 |
| 2 | 1 | 0 | 0 | 1 | 0 | 0 |
| 3 | 0 | 1 | 0 | 0 | 1 | 0 |
| 4 | 0 | 0 | 1 | 0 | 0 | 1 |

حداقل تعداد ستون برای یک ویژگی با n مقدار n-1 است.( با دانستن وضعیت n-1 مقدار وضعیت مقدار نهایی مشخص است)

ج)

Hot-One Encoding:

- مناسب برای متغیرهای دسته‌ای با مقادیر گسسته و ترتیبی نیستند. به عبارت دیگر، اگر وجود ترتیب بین مقادیر دسته‌ای مهم نباشد و تنها مهم باشد که مقادیر متفاوت باشند، این روش مناسب است.
- معمولاً برای متغیرهایی با تعداد مقادیر کمتر مناسب است، زیرا طول بردارهای باینری برابر تعداد مقادیر ممکن است و در صورت وجود تعداد زیادی از مقادیر، این روش می‌تواند به وجود بردارهای بسیار بزرگ منجر شود که ممکن است باعث افزایش پیچیدگی و حافظه مصرفی شود.

Label Encoding:

- مناسب برای متغیرهای دسته‌ای با مقادیر گسسته و ترتیبی هستند. در این روش، مقادیر دسته‌ای به صورت ترتیبی به اعداد صحیح نگاشت می‌شوند.
- معمولاً برای متغیرهایی با تعداد مقادیر زیاد مناسب است، زیرا در Label Encoding تنها تک عدد صحیح به هر مقدار اختصاص می‌یابد و به همین دلیل از لحاظ حافظه و پیچیدگی کمتری نسبت به Hot-One

**سوال ۵.**

درکاهش بعد دادهها روی یک بعد، PCAتلاش می کند خطی را برای تصویرکردن داده ها انتخاب کند که تصویر داده ها روی آن خط بیشترین واریانس را داشته باشد یا به طور معادل تا حد ممکن فاصله بین نقاط از هم و فاصله تصویر آنها روی خط به یکدیگر نزدیک باشد. در اینجا مجموعه نقاط دادهشده، همگی ضریبی از $\boldsymbol{v}$ هستند، بنابراین روی یک خط قرار دارند. پس خط مطلوب، خطی است که در راستای $\boldsymbol{v}$ بوده و از مبدأ می گذرد، فرم پارامتری معادله این خط بهصورت زیر می باشد:

$$\vec{r} = t\vec{v}, \quad t \in \mathbb{R}$$

در اینصورت تصویر هر نقطه خودش می باشد و بنابراین فاصله هر دو نقطه از هم با فاصله تصویر آن دو نقطه روی خط برابر است و ما به خط مطلوب دست یافته ایم. برای یافتن مؤلفه اصلی نیز کافی است نرمالشده بردار $\boldsymbol{v}$ را درنظر بگیریم:

$$\boldsymbol{u} = \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|}$$

تصویر هر نقطه روی این خط نیز خود نقطه میباشد که اگر طبق ضابطه PCA هم این نقاط را به دست آوریم، خواهیم داشت:

$$\boldsymbol{u}^T\boldsymbol{v} = \frac{\boldsymbol{v}^T}{\|\boldsymbol{v}\|}\boldsymbol{v} = \frac{\|\boldsymbol{v}\|^2}{\|\boldsymbol{v}\|} = \|\boldsymbol{v}\|,$$
$$\boldsymbol{u}^T(-\boldsymbol{v}) = -\frac{\boldsymbol{v}^T}{\|\boldsymbol{v}\|}\boldsymbol{v} = -\frac{\|\boldsymbol{v}\|^2}{\|\boldsymbol{v}\|} = -\|\boldsymbol{v}\|,$$
$$\boldsymbol{u}^T(2\boldsymbol{v}) = \frac{\boldsymbol{v}^T}{\|\boldsymbol{v}\|}(2\boldsymbol{v}) = 2\frac{\|\boldsymbol{v}\|^2}{\|\boldsymbol{v}\|} = 2\|\boldsymbol{v}\|,$$
$$\boldsymbol{u}^T(-2\boldsymbol{v}) = -2\frac{\boldsymbol{v}^T}{\|\boldsymbol{v}\|}\boldsymbol{v} = -2\frac{\|\boldsymbol{v}\|^2}{\|\boldsymbol{v}\|} = -2\|\boldsymbol{v}\|,$$

**سوال ۶.**

طبق تعریف مسئله داریم:

$$\beta_0 = \bar{y} - \beta_1 \bar{x} \qquad \beta_1 = \frac{\sum_{i=0}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=0}^{n}(x_i - \bar{x})^2}$$

همچنین می‌دانیم:

$$cov(\alpha x + \beta y, \gamma w + \delta z) = \alpha\gamma\, cov(x, w) + \alpha\delta\, cov(x, z) + \beta\gamma\, cov(y, w) + \beta\delta\, cov(y, z) \quad (۴)$$

$$cov(x, \alpha) = 0 \quad (۵)$$

حال می‌خواهیم کوواریاس مربوط به دو پارامتر را حساب کنیم:

$$cov(\beta_0, \beta_1) = cov(\bar{y} - \beta_1 \bar{x}, \beta_1)$$

با توجه به ۴ و اینکه $\overline{x}$ و $\overline{y}$ مقادیر ثابتی هستند داریم:

$$\begin{aligned} cov(\beta_0, \beta_1) &= cov(\bar{y}, \beta_1) - \bar{x}\, cov(\beta_1, \beta_1) = 0 - \bar{x}\, var(\beta_1) \\ &= -\bar{x}\, var(\beta_1) \\ &= -\bar{x}\, var\Big(\frac{\sum_{i=0}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=0}^{n}(x_i - \bar{x})^2}\Big) \\ &= -\bar{x}\, var\Big(\frac{\sum_{i=0}^{n}(y_i(x_i - \bar{x}) - \bar{y}(x_i - \bar{x}))}{\sum_{i=0}^{n}(x_i - \bar{x})^2}\Big) \\ &= -\bar{x}\, var\Big(\frac{\sum_{i=0}^{n} y_i(x_i - \bar{x}) - \bar{y}\sum_{i=0}^{n}(x_i - \bar{x})}{\sum_{i=0}^{n}(x_i - \bar{x})^2}\Big) \end{aligned}$$

اما باید توجه داشته باشیم که:

$$\sum_{i=0}^{n}(x_i-\bar{x})=\sum_{i=0}^{n}x_i-\sum_{i=0}^{n}\bar{x}=\sum_{i=0}^{n}x_i-n\bar{x}=0\rightarrow\bar{x}=\frac{\sum_{i=0}^{n}x_i}{n}$$

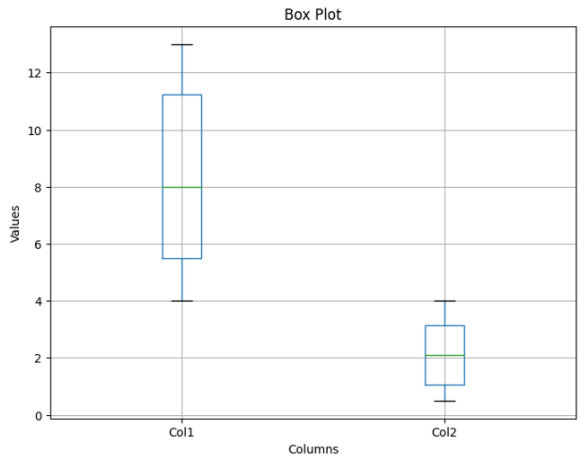
$$\begin{aligned}cov(\beta_0,\beta_1)&=-\bar{x}var\Big(\frac{\sum_{i=0}^{n}y_i(x_i-\bar{x})}{\sum_{i=0}^{n}(x_i-\bar{x})^2}\Big)\\&=-\bar{x}\Big(\frac{\sum_{i=0}^{n}var(y_i)(x_i-\bar{x})^2}{(\sum_{i=0}^{n}(x_i-\bar{x})^2)^2}\Big)\\&=-\bar{x}\Big(\frac{\sigma^2\sum_{i=0}^{n}(x_i-\bar{x})^2}{(\sum_{i=0}^{n}(x_i-\bar{x})^2)^2}\Big)\\&=-\bar{x}\Big(\frac{\sigma^2}{\sum_{i=0}^{n}(x_i-\bar{x})^2}\Big)\end{aligned}$$

**سوال ۷.**

الف)

نمودار Box plot یک نمودار توصیفی است که توزیع داده‌های عددی را به صورت مرتب و قابل مقایسه نمایش می‌دهد. این نمودار شامل خط مرکزی که نشان‌دهندهٔ میانه است، جعبه که حاوی کوارتیل‌ها است، دم‌ها که محدودهٔ احتمالی داده‌ها را نشان می‌دهند، و ابرصورت‌ها که نقاط پرت را نشان می‌دهند، می‌باشد.

کاربردهای نمودار Box plot عبارتند از: مقایسهٔ توزیع داده‌ها در گروه‌های مختلف، تشخیص داده‌های پرت و نقاط نامتعادل، تحلیل توزیع داده‌ها و مدیریت داده‌های پرت. این نمودار به محققان و تحلیلگران در زمینه‌های مختلف کمک می‌کند تا به طور سریع و دقیق ویژگی‌های مهم توزیع داده‌ها را درک کنند و تفاوت‌ها و الگوهای آماری را در داده‌ها مشاهده کنند.



**سوال ۸.**

گسترش هر یک از ترمها سوال را حل می‌کنیم.

بخش اول و دوم شبیه هم گسترش میدیم:

$$||\overline{X_i}||^2 + \sum_{p=1}^{n} ||\overline{X_p}||^2/n.$$

$$||\overline{X_j}||^2 + \sum_{q=1}^{n} ||\overline{X_q}||^2/n.$$

بخش سوم به صورت زیر:

$$\sum_{p=1}^{n} ||\overline{X_p}||^2/n + \sum_{q=1}^{n} ||\overline{X_q}||^2/n$$

پاسخ سوال نهم:

**سوال ۹.**

برای حل این سوال انتگرال زیر را حل می کنیم که با توجه به این که متغییر ها جدایی پذیر است و تابع ما نمایی است و در مواجهه با log به صورت خطی ساده می شود نیاز به حل انتگرال پیچیده و دو گانه نیست منتها دقت شود که لگاریتم گرفته شده در مبنای ۲ است و باید تغییر مبنا صورت گیرد

$$I(x,y) = \iint_{0}^{\infty} e^{-(x+y)} \log(e^{-(x+y)}) dxdy = 2.88\ bit$$