دانشگاه صنعتی اصفهان
دانشکده مهندسی برق و کامپیوتر

داده کاوی

تمرین شماره ۲

زمستان ۱۴۰۲

فهرست مطالب

۱ سوالات

۱.۱ سوال ۱

(آ) شاخصه های آماری(میانگین، میانه، مد، چارک و واریانس) را برای مجموعه داده زیر که نشان دهنده درصد چربی در یک نمونه از محصولات غذایی است را به صورت دستی محاسبه کنید:
داده ها: ۲۱، ۱۵، ۱۸، ۲۳، ۲۰، ۱۷، ۱۶، ۱۹، ۲۲، ۲۴، ۲۷، ۱۸، ۲۰

(ب) با استفاده از توابع آماده در کتابخانه های پایتون نظیر numpy و ... شاخصه های آماری را برای بخش قبلی در پایتون محاسبه کنید.

(ج) با استفاده کتابخانه matplotlib یا سایر کتابخانه های مشابه در پایتون، boxplot را برای این داده ها رسم کنید.

(د) مقدار Z-Score را برای این داده ها به صورت دستی محسابه کنید.

(ه) با استفاده از کتابخانه های پایتون مقدار Z-Score را برای این داده ها به دست آورید.

۲.۱ سوال ۲

برای موارد زیر، مشخص کنید چه نوع داده ای داریم ؟ (جدولی، گراف، Ordered)

(آ) داده های سری زمانی قیمت سهام برای یک شرکت معین در طول یک ماه.

(ب) داده های موجودی در یک سیستم مدیریت انبار

(ج) آمار ورزشی برای ورزشکاران

(د) شبکه های حمل و نقل که مسیرها و اتصالات بین فرودگاه ها را نشان می دهند.

(ه) داده های بیان ژن در مراحل مختلف رشد در یک ارگانیسم

(و) شبکه‌های جاده‌ای که خیابان‌ها و تقاطع‌ها را در نقشه شهر نشان می‌دهند.

(ز) توپولوژی اینترنت که اتصالات بین روترها و شبکه ها را نشان می دهد.

(ح) جدول زمانی تکامل گونه ها بر اساس سوابق فسیلی.

(ط) اطلاعات مشتری در یک سیستم CRM

(ی) داده های گزارش متوالی که تعاملات کاربر را در یک وب سایت ثبت می کند.

(ک) شبکه ای از تعاملات بین پروتئین ها در یک سیستم بیولوژیکی

(ل) نمرات دانش آموز در یک سیستم مدرسه

(م) روابط هم نویسندگی بین محققان در یک شبکه انتشارات علمی

(ن) روابط وابستگی بین ماژول های نرم افزار در یک پایگاه کد.

(س) خواندن علائم حیاتی بیمار در فواصل منظم در بیمارستان.

۳.۱ سوال ۳

فاصله Mahalanobis را برای مقادیر زیر به دست آورید:

values:

X=5 , Y=8

Mean vector (μ) = (4.7)

Covariance matrix (Σ):

110 41

14 51

۴.۱ سوال ۴

مقدار SMC و Jaccard را برای مجموعه های زیر حساب کنید:

A = {apple, banana, cherry, date}

B = {banana, cherry, date, elderberry}

۵.۱ سوال ۵

فرض کنید یک متغیر تصادفی X با سه نتیجه ممکن داریم: A ، B و C ، با احتمالات زیر:

P(A) = 0.4

P(B) = 0.3

P(C) = 0.3

آنتروپی را برای آن محسابه کنید.

۶.۱ سوال ۶

فایل LaptopSalesJanuary2008.csv حاوی داده‌هایی برای تمام فروش لپ‌تاپ‌ها در یک زنجیره کامپیوتر در لندن در ژانویه ۲۰۰۸ است. این زیرمجموعه‌ای از مجموعه داده کامل است که شامل داده‌های کل سال است.

۱. با استفاده از کتابخانه های پایتون، یک نمودار میله ای ایجاد کنید که میانگین قیمت خرده فروشی را بر اساس فروشگاه نشان دهد. کدام فروشگاه بالاترین میانگین را دارد؟ کدام کمترین را دارد؟

۲. برای مقایسه بهتر قیمت‌های خرده‌فروشی در فروشگاه‌ها، نمودار هایی جعبه ای از قیمت خرده‌فروشی به فروشگاه ایجاد کنید. اکنون قیمت های موجود در دو فروشگاه (به دست آمده در بخش آ) را مقایسه کنید.
آیا تفاوتی بین توزیع قیمت آنها وجود دارد؟ توضیح دهید.

۷.۱ سوال ۷

در این سوال با استفاده از داده هایی که با همکاری دوستانتان در تمرین شماره یک جمع آوری شد، میخواهیم به تحلیل برخی از بیماری ها (با تمرکز بر مصور سازی داده ها) بپردازیم.
مراحل زیر را به صورت گام به گام دنبال کنید:

۱. از منابع مختلف در مورد ساختار فایل CSV در داده کاوی و مزایای آن تحقیق کنید و توضیح دهید. (به صورت مختصر)

۲. یک فایل CSV با یک ویرایشگر متنی ساده نظیر Notepad در ویندوز یا gedit در لینوکس ایجاد کنید(ایجاد فایل CSV در اکسل امکان پذیر است اما برای بارگذاری آن در پایتون ممکن است با مشکلاتی مواجه شوید که البته راهکارهای خاص خود را دارند.)
سپس اطلاعات حداقل ۳ بیماری دلخواه را در ستون های گسسته به روش زیر ثبت کنید. (به دلیل جلوگیری از برخی پیچیدگی ها لازم است تنها نام بیماری، توصیه های بیماری و شرایط مراجعه به پزشک را درنظر داشته باشید.)

1.csv

```
Name, History of illness in the family, Chest pain, high blood sugar, sports activity , Diet control, Use of an insulin pump
Type 1 diabetes,1,0,1,1,1,1
blood fat,1,1,1,1,1,0
```

در مثال فوق برای دو بیماری چربی خون و دیابت برای هر (علامت جهت مراجعه به پزشک) و هر (توصیه) در مقابل بیماری یک ستون در نظر گرفته شده و با مقدار ۰ یا ۱ رکورد ها مشخص می شوند. (تعداد ستون های شما طبیعتا بیشتر خواهد بود.)

۳. پس از لود کردن داده ها در Notebook خود ، با استفاده از کتابخانه های مصورسازی داده ها سعی کنید بیماری هایی با شرایط مشترک مراجعه به پزشک و توصیه های مشترک را پیدا کنید. از حداقل ۳ مدل نمودار برای این سوال کمک بگیرید.

نمودار های رایج: (Line Plot , Bar Plot , Histogram , Scatter Plot , Pie Chart, Box Plot , Heatmap , Violin Plot, Area Plot)

نکته: تاکیدی بر استفاده از کتابخانه خاصی در این سوال مطرح نیست و با کتابخانه های دلخواه نظیر seaborn, matplotlib و ... میتوانید برای حل این مسئله تلاش کنید.

نمونه کد برای مثال مطرح شده(seaborn):

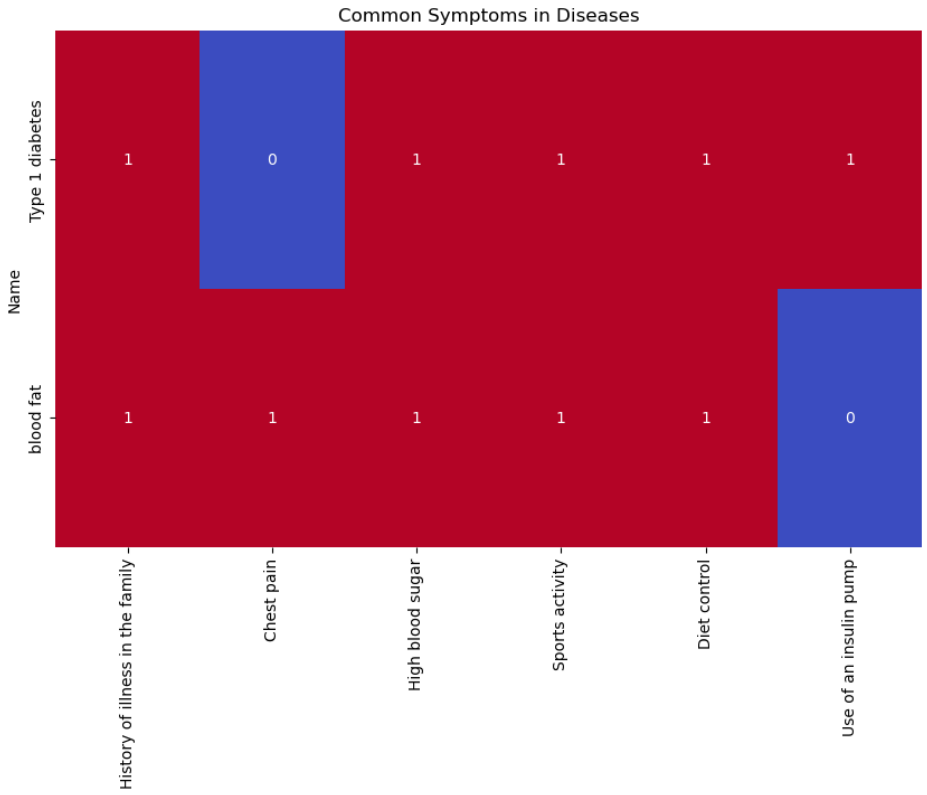
```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

df = pd.read_csv('1.csv')
```

```
# Set the index
df.set_index('Name', inplace=True)

# Create a heatmap
plt.figure(figsize=(10, 6))
sns.heatmap(df, cmap='coolwarm', annot=True, cbar=False)
plt.title('Common Symptoms in Diseases')
plt.show()
```

خروجی :



۲ نکات پاسخ دهی

- راهنمای نصب محیط برنامه نویسی برای این تمرینات به همراه برخی نکات مورد نیاز برای انجام تمرینات به صورت ویدئو در سامانه یکتا بارگزاری شده است.
- با جستجو در سایت های مختلف می توانید به کدهای مشابه برای انجام این تمرینات دسترسی داشته باشید.
- برای تمرینات غیر عملی که به صورت تایپی ارسال شوند امتیاز تشویقی درنظر گرفته می‌شود.
- فایل پایتون و یا Notebook برای تمرینات ضمیمه شود و همه به صورت یک فایل zip بارگذاری شوند. فایل zip را با فرمت DM4022_HW2_[StudentNumber].zip نام‌گذاری کنید.
- درصورت وجود ابهام خاص می توانید موارد را با دستیار آموزشی مطرح کنید.
  ایمیل: q.soleimani@ec.iut.ac.ir