

به نام خدا



دانشگاه تهران



دانشکده مهندسی برق و کامپیوتر

درس داده کاوی تمرین اول

محمدجواد کامیاب mj.kamyab@ut.ac.ir	طراحان
۱۴۰۲/۱۲/۰۸	تاریخ بارگذاری
۱۴۰۲/۱۲/۲۳	مهلت ارسال

فهرست

۳.....	بخش تشریحی
۳.....	سوال اول
۴.....	سؤال دوم
۵.....	بخش عملی
۵.....	پیش‌نیازها
۶.....	شرح دادگان
۷.....	پیش‌پردازش
۸.....	نمایش دادگان
۱۱.....	ملاحظات

جدول‌ها

جدول ۱. توضیحات مجموعه دادگان..... ۶

سوال اول

داده‌های زیر را در نظر بگیرید و برای هر کدام مشخص کنید که از چه نوعی است (پیوسته، گسسته، باینری، اسمی یا ترتیبی^۱):

- سن
- جنسیت
- میزان درآمد
- وضعیت تاهل
- فرزند دارد
- شغل
- میزان تحصیلات
- تعداد اعضای خانواده

هر کدام از ویژگی‌ها بالا را با چه نمودارهایی می‌توان نمایش داد؟

Histogram, Pie chart, Box plot, Bar chart

ممکن است بعضی از ویژگی‌ها را با چند نمودار بتوان نمایش داد.

¹ Continuous, Discrete, Binary, Nominal, or Ordinal

سؤال دوم

دو سری ویژگی زیر را در نظر بگیرید:

$$A = \{55, 72, 60, 54, 42, 64, 43, 89, 96, 38, 79, 52, 56, 92, \\ 7, 8, 24, 39, 44, 68, 68, 52, 4, 16, 73, 46, 96, 38, 20, 27\}$$
$$B = \{11, 16, 13, 11, 9, 14, 9, 19, 20, 8, 17, 11, 12, 20, \\ 2, 3, 5, 8, 9, 14, 14, 11, 2, 4, 15, 9, 21, 8, 4, 5\}$$

۱. میانگین، میانه، چارک اول، چارک سوم و انحراف معیار را برای هر کدام از خصیصه‌ها محاسبه کنید.
۲. نمودار جعبه‌ای برای دو خصیصه را رسم کنید و میزان پراکندگی داده‌ها در این دو خصیصه را مقایسه کنید.
۳. نمودار هیستوگرام را برای دو خصیصه را رسم کنید.
۴. ابتدا هر خصیصه را با استفاده از z-score نرمال کنید و سپس توزیع مقادیر دو خصیصه A و B را با رسم نمودار plot Q-Q مقایسه کنید.
۵. آیا این دو خصیصه با هم هم‌بستگی^۱ دارند؟ توضیح دهید.

نکته: برای این بخش نیاز به محاسبه‌ی تمامی مقادیر به صورت دستی نیست تنها فرمول مورد استفاده را نوشته و بعد از انجام یک مورد برای محاسبه‌ی سایر موارد می‌توانید از Excel استفاده کنید. همچنین برای رسم نمودارها می‌توانید از Excel یا هر ابزاری که مد نظر دارید اسفاده کنید.

^۱ correlation

پیش‌نیازها

برای پاسخ به این تمرین عملی باید از زبان برنامه‌نویسی **Python** استفاده کنید و نیاز است که پیش از شروع، یک سرور **Jupyter** بر روی سیستم نصب و راه‌اندازی شود تا بتوانید بر روی یک فایل **.ipynb** کدهای خود را اجرا کنید، همچنین راه حل جایگزین آن استفاده از **Google Colab** است.

استفاده از کتابخانه‌های **Pandas**، **Numpy** و **Datetime** می‌تواند گزینه‌ی مناسبی برای حل مسائل پیشرو باشد، همچنین دو کتابخانه‌ی **Matplotlib** و **Seaborn** در بخش مصورسازی مجموعه داده‌گان بسیار مفید هستند.

شرح دادگان

این مجموعه داده با نام weatherAUS در فایل فشرده dataset.zip قرار داده شده و شامل اطلاعات مربوط به آب و هوای استرالیا است. این مجموعه داده شامل ۲۳ ستون است و اطلاعات مربوط به هر ستون در جدول زیر آورده شده است.

جدول ۱. توضیحات مجموعه دادگان

#	Column name	Description
1	Date	The date of observation
2	Location	The common name of the location of the weather station
3	MinTemp	The minimum temperature in degrees Celsius or Fahrenheit
4	MaxTemp	The maximum temperature in degrees Celsius or Fahrenheit
5	Rainfall	The amount of rainfall recorded for the day in mm
6	Evaporation	The so-called Class A pan evaporation (mm) in the 24 hours to 9am
7	Sunshine	The number of hours of bright sunshine in the day.
8	WindGustDir	The direction of the strongest wind gust in the 24 hours to midnight
9	WindGustSpeed	The speed (km/h) of the strongest wind gust in the 24 hours to midnight
10	WindDir9am	Direction of the wind at 9am
11	WindDir3pm	Wind speed (km/hr) averaged over 10 minutes prior to 9am
12	WindSpeed9am	Wind speed (km/hr) averaged over 10 minutes prior to 9am
13	WindSpeed3pm	Wind speed (km/hr) averaged over 10 minutes prior to 3pm
14	Humidity9am	Humidity (percent) at 9am
15	Humidity3pm	Humidity (percent) at 3pm
16	Pressure9am	Atmospheric pressure (hpa) reduced to mean sea level at 9am
17	Pressure3pm	Atmospheric pressure (hpa) reduced to mean sea level at 3pm
18	Cloud9am	Fraction of sky obscured by cloud at 9am. This is measured in "oktas", which are a unit of eighths. It records how many eighths of the sky are obscured by cloud. A 0 measure indicates completely clear sky whilst an 8 indicates that it is completely overcast.
19	Cloud3pm	Fraction of sky obscured by cloud (in "oktas": eighths) at 3pm. See Cloud9am for a description of the values
20	Temp9am	Temperature (degrees C) at 9am
21	Temp3pm	Temperature (degrees C) at 3pm
22	RainToday	Boolean: 1 if precipitation (mm) in the 24 hours to 9am exceeds 1mm, otherwise 0
23	RainTomorrow	The amount of next day rain in mm. Used to create response variable RainTomorrow. A kind of measure of the "risk".

پیش‌پردازش

پیش‌پردازش، یکی از مهم‌ترین گام‌ها در پروژه‌های داده‌کاوی است. رویکردهای مختلفی در زمینه‌ی مدیریت داده‌های گم شده و تبدیل داده‌ها به فرمت‌های دیگر مورد استفاده قرار می‌گیرد و انتخاب دقیق این رویکردها تأثیر مستقیمی در کیفیت نتایج نهایی دارد؛ لذا همواره می‌بایست بهترین رویکرد را شناسایی و اعمال نمود.

۱. این دیتاست از تعداد زیادی ایستگاه هواشناسی در شهرهای مختلف جمع‌آوری شده‌است. ابتدا تمامی فایل‌ها را با یکدیگر ادغام کنید. در حین ادغام ممکن است به مشکلاتی برخورد کنید، به دلیل اینکه فایل‌ها از ایستگاه‌های متفاوتی جمع‌آوری شده نام ستون‌ها در این فایل‌ها ممکن است اندکی متفاوت باشد یا بعضی از ایستگاه‌ها برخی از پارامترها را محاسبه نکرده باشند، سایر مشکلاتی که به آن برخورد کردید را ذکر کنید.
۲. ابتدا ۵ سطر ابتدایی دیتاست را نمایش دهید.
۳. نشان دهید هر ستون با چه نوع متغیری ذخیره شده‌است. (object, int64, string, ...)
۴. با توجه به اینکه دقت اعشاری برای دما نیاز نداریم تمامی ستون‌هایی که دما را مشخص می‌کنند، باید به یک عدد صحیح گرد شوند.
۵. حجمی که دیتاست در حافظه RAM اشغال کرده است چقدر است؟
۶. با تغییر نوع متغیرها حجم این مجموعه داده را کمتر کنید. به عنوان مثال: مسیر باد که یک داده‌ی اسمی است اگر به نوع متغیر category تبدیل شود بسیار حجم را کاهش می‌دهد، علت این تغییر حجم را ذکر کنید، چه تفاوتی در نوع ذخیره سازی قبل و بعد از این تبدیل متغیر وجود دارد که باعث این کاهش حجم شده‌است. بهترین نوع متغیر را برای هر کدام از ستون‌ها گزارش کنید. (برای این بخش بهتر است از تابع astype که در pandas وجود دارد استفاده کنید)
۷. پس از تبدیلات ذکر شده حجم دیتاست چه مقدار و چند درصد کاهش یافت؟
۸. برای هر کدام از ستون‌های این مجموعه داده تعداد مقادیر گم‌شده^۱ را گزارش کنید.
۹. چه رویکردی برای هر کدام از ستون‌ها مناسب است دلیل آن را ذکر کنید. (راه‌حل‌ها می‌تواند شامل حذف ردیف‌های شامل داده گم‌شده یا درج کردن داده در آن مکان باشد)
۱۰. با استفاده از روش‌های انتخاب شده در سؤال ۹ مشکل داده‌های گم‌شده را برطرف کنید.
۱۱. مقدار ذخیره شده دما در شهرهای مختلف با واحدهای اندازه‌گیری متفاوتی ثبت شده‌است (بعضی سلسیوس و برخی دیگر فارنهایت). همه را به سلسیوس تبدیل کنید.
۱۲. در این قسمت داده‌های پرت^۲ را شناسایی کنید و روش برخورد را بیان کنید.

¹ Missing value

² Outliers

نمایش دادگان

مصورسازی داده‌ها با استفاده از نمودارهای مناسب دید بهتری نسبت به اطلاعات موجود در مجموعه داده را ایجاد می‌کند و همچنین می‌تواند باعث شود تحلیل‌گران و مدیران تصمیمات بهتری اتخاذ کنند.

• تمامی نمودارها با عنوان، اندازه و نام مناسب برای هر محور رسم شود.

۲. میانگین دما در ایستگاه را با نمودار میله‌ای^۱ نمایش دهید (برای هر روز ابتدا میانگین دمای آن روز را نیز باید محاسبه کنید، میانگین حداقل و حداکثر دما) و همچنین مقدار میانگین در نمودار بر روی هر میله با فونت و رنگ مناسب نوشته شده باشد. (برای نمایش داده‌ها با استفاده از نمودار میله‌ای زمانی که تعداد میله‌ها زیاد باشد، بهتر است که میله‌ها به صورت افقی باشند، همچنین مرتب کردن مقدارها به صورت نزولی، دید بهتری را به بیننده می‌دهد)

۳. پنج ایستگاه که سریع‌ترین بادها را ثبت کرده‌اند را با استفاده از نمودار مناسب نمایش دهید.

۴. میزان تابش نور خورشید در هر روز چه تأثیری بر روی دما دارد؟

۵. بررسی کنید کدام یک از متغیرها با یکدیگر ارتباط بیشتری دارند و این ارتباط را با نمودار مناسب نمایش دهید (از معیار Correlation استفاده کنید، نمودار heatmap نیز در بخش می‌تواند برای نمایش زوج ارتباطات مناسب باشد)

۶. میزان بارش را با نمودارهای جعبه‌ای^۲، فراوانی^۳ و توزیعی^۴ نمایش دهید. کدام نمودار اطلاعات بیشتری را به بیننده می‌دهد (بعد از رسم نمودار جعبه‌ای اگر نمودار به اندازه‌ی کافی اطلاعات در اختیار نمی‌گذارد بهتر است که محدوده‌ی ورودی را محدود کنیم، مثلاً ۵ درصد بیشترین و کمترین دماها را حذف کنیم تا نمودار بدون داده‌ی پرت نمای بهتری داشته باشد همچنین از Scale کردن محورها نیز می‌توانید استفاده کنید).

۷. با استفاده از نمودار پراکندگی^۵ می‌توانیم ارتباط دو متغیر با یکدیگر در فضای دو بعدی را ترسیم کنیم، آیا نمودار دیگری یا حالت خاصی از این نمودار وجود دارد که بتواند در فضای دو بعدی ارتباط بیشتر از دو متغیر را نمایش دهد؟ در صورت وجود، این نمودار را برای بیش از دو پارامتر به شکل معناداری ترسیم کنید.

¹ Bar plot

² Box plot

³ Histogram

⁴ Distribution plot

⁵ Scatter plot

۸. یک ایستگاه هواشناسی را انتخاب کنید و نمودارها زیر را برای این ایستگاه رسم کنید:

- نمودار میزان بارش در روزهای مختلف
- نمودار میزان تبخیر در روزهای متفاوت
- نمودار میزان سرعت باد در روزهای متفاوت


۹. تمام دماهای ثبت شده در تمامی ایستگاهها را در نظر بگیرید، این دماها را به پنج دسته تقسیم کنید:

- quartile 0-10 روز خیلی سرد
- quartile 10-30 روز سرد
- quartile 30-70 روز معمولی
- quartile 70-90 روز گرم
- quartile 90-100 روز خیلی گرم

حال ایستگاههایی که بیشترین روز در هر دسته بندی داشته اند را گزارش کنید.

۱۰. (امتیازی) می‌خواهیم برای هر ایستگاه میانه‌ی ستون‌های حداکثر دما، حداقل دما، دما در ساعت ۹ صبح و دما در ساعت ۳ بعد از ظهر را محاسبه کنیم و نمایش دهیم. برای این کار ابتدا میانه را محاسبه می‌کنیم، بعد از این مرحله به یک دیتا فریم با پنج ستون خواهیم رسید که اولین ستون نام ایستگاه‌ها و ستون‌های دیگر دما در حالت‌ها متفاوت است؛ حال از دستور melt از کتابخانه‌ی pandas استفاده می‌کنیم تا تمامی مقادیر در یک ستون قرار گیرد و اسامی این مقادیر نیز در یک ستون دیگر قرار گیرد به‌عنوان مثال:

Location	MinTemp	MaxTemp	Temp9am	Temp3pm
Adelaide	12.0	21.0	16.0	20.0
Albany	12.0	19.0	16.0	17.0
Albury	9.0	21.0	14.0	20.0
AliceSprings	13.0	29.0	21.0	28.0
BadgerysCreek	11.0	23.0	16.0	21.0
Ballarat	7.0	16.0	10.0	15.0
Bendigo	8.0	20.0	13.0	19.0
Brisbane	16.0	26.0	22.0	24.0
Cairns	21.0	29.0	26.0	28.0
Canberra	7.0	20.0	12.0	18.0
Cobar	13.0	25.0	18.0	24.0



Location	variable	value
Adelaide	MinTemp	12.0
Adelaide	MaxTemp	21.0
Adelaide	Temp3pm	20.0
Adelaide	Temp9am	16.0
Albany	Temp3pm	17.0
Albany	MinTemp	12.0
Albany	MaxTemp	19.0
Albany	Temp9am	16.0
Albury	MaxTemp	21.0
Albury	Temp9am	14.0
Albury	MinTemp	9.0
Albury	Temp3pm	20.0
AliceSprings	MaxTemp	29.0
AliceSprings	Temp9am	21.0
AliceSprings	Temp3pm	28.0
AliceSprings	MinTemp	13.0

با این تغییرات نمودار میله‌ای را با استفاده از کتابخانه‌ی seaborn رسم کنید که محور y اسم هر ایستگاه، محور x میزان دما و مقدار hue برابر نوع متغیر می‌شود.

- هدف این بخش آشنایی با تابع melt است در مورد این تابع و روش استفاده از این تابع توضیح دهید همچنین یک تابع عمل مخالف melt را انجام می‌دهد، آن را نیز معرفی و کاربرد آن را تشریح کنید.

ملاحظات

- تمامی نتایج شما باید در یک فایل فشرده با عنوان DM_CA1_StudentID تحویل داده شود.
- خوانایی و دقت بررسی‌ها در گزارش نهایی از اهمیت ویژه‌ای برخوردار است. به تمرین‌هایی که به صورت کاغذی تحویل داده شوند یا به صورت عکس در سایت بارگذاری شوند، ترتیب اثری داده نخواهد شد.
- بخش اصلی نمره به گزارش شما تعلق می‌گیرد و دستیاران الزامی برای اجرای تمام کدهای شما در صورتی که در گزارش به آن‌ها اشاره‌ای نکرده باشید ندارند. لطفاً تمام موارد مورد نیاز را در گزارش ذکر کنید.
- کدهای نوشته شده برای هر بخش را با نام مناسب مشخص کرده و به همراه گزارش تکلیف ارسال کنید. همه‌ی کدهای پیوست گزارش بایستی قابلیت اجرای مجدد داشته باشند. در صورتی که برای اجرا مجدد آنها نیاز به تنظیمات خاصی می‌باشد بایستی تنظیمات مورد نیاز را نیز در گزارش خود ذکر کنید.
- برای تحویل تمارین از چارچوب قرارداده شده در سامانه و کانال تلگرام استفاده کنید.
- در صورت قصد ارسال تمرین به صورت دیگر (انگلیسی، latex و ...)، لطفاً پیش از ارسال با دستیارمسئول تمرین هماهنگ کنید.
- توجه کنید این تمرین باید به صورت تک نفره انجام شود و پاسخ‌های ارئه شده باید نتیجه فعالیت فرد نویسنده باشد (هم‌فکری خارج از چارچوب و به اتفاق هم نوشتن تمرین نیز ممنوع است). در صورت مشاهده تقلب برای همه‌ی افراد مشارکت کننده، نمره تمرین، صفر در نظر گرفته خواهد شد.
- در صورت بروز هرگونه مشکل با ایمیل زیر در ارتباط باشید:

<mailto:mj.kamyab@ut.ac.ir>

مهلت تحویل: ۲۳ اسفند ۱۴۰۲

مهلت تحویل با تاخیر: ۱ فروردین ۱۴۰۳