

تمرین اول درس مبانی داده کاوی  
(زمستان ۴۰۱)

مهلت تحویل تمرین: ۱۷ اسفند ماه

سوالات تئوری

سوال ۱- فرض کنید داده هایی در مورد مراجعان یک بیمارستان در دسترس است. این داده ها می تواند شامل سن، جنسیت، سابقه بیماری قبلی، شغل، قد، وزن و ... باشد. با در نظر گیری داده های موجود چهار نمونه مسئله با استفاده از تسک های داده کاوی مانند پیش بینی، دسته بندی و ... معرفی و توصیف نمایید.

سوال ۲- صفات زیر را در دسته های ارائه شده طبقه بندی کنید. در صورت ابهام با توضیح دلیل انتخاب خود را بیان کنید.

- Binary or not
  - Discrete or Continuous
  - qualitative (nominal or ordinal) or quantitative (interval or ratio)
- سن بر حسب سال
  - روشنایی که با نورسنج اندازه گیری می شود
  - روشنایی که با نظر افراد بیان می شود
  - زاویه اندازه گیری شده با وسیله اندازه گیری (نقاله و ...)
  - مدال های اهدایی در مسابقات المپیک
  - ارتفاع از سطح دریا
  - تعداد بیماران یک بیمارستان
  - شماره ISBN (در مورد این شماره و نحوه اختصاص آن به کتاب در اینترنت جستجو کنید).
- سوال ۳ - جدول زیر مقادیر ثبت شده برای قیمت کالاهای استوک وارداتی یک شرکت را نشان می دهد.

10	7	20	12	75	15	9	18	4	12	8	14
----	---	----	----	----	----	---	----	---	----	---	----

- لطفا مقادیر زیر را محاسبه کنید:
- ۱: میانگین
- ۲: میانه
- ۳: مد
- ۴: انحراف معیار
- ۵: شاخص zscore

سوال ۴ (اختیاری) - تمرین ۲.۶ (سوال ۶ فصل دوم) کتاب آقای هان را حل نمایید. (Data Mining\_ Concepts and Techniques)

## سوالات عملی

سوال ۵ - از دیتاست Smartphone که در اختیارتان قرار داده شده است برای حل سوالات زیر استفاده نمایید:

اطلاعات بیشتر درباره ستون های دیتاست را می توانید در این [لینک](#) جستجو کنید.

- ۱- نخست دیتاست را با استفاده از کتابخانه pandas خوانده و تبدیل به دیتافریم نمایید.
- ۲- اطلاعات توصیفی دیتاست مانند تعداد و نوع داده های هر ستون و حجم دیتاست را نمایش دهید.
- ۳- سطرهایی از دیتاست که مربوط به تلفن هایی است که قابلیت اتصال به بلوتوث دارد و هم حجم حافظه داخلی آن از ۱۰ گیگابایت بیشتر است را جداسازی کرده و نمایش دهید.
- ۴- اطلاعات تلفن هایی که دارای قوی ترین و ضعیف ترین دوربین جلو هستند را از دیتاست استخراج کرده و نمایش دهید.
- ۵- ستون جدیدی به نام clock\_rate به دیتاست اضافه نمایید و به ازای تلفن هایی که clock\_speed نظیر آن ها برابر و یا کوچکتر از ۱ است مقدار این ستون را "Low"، به ازای آن هایی که از ۱ بزرگتر و از ۲ کوچکتر است مقدار این ستون را "Mid" و به ازای آن هایی که clock\_speed بیش از ۲ دارند مقدار ستون جدید را برابر "High" قرار دهید.
- ۶- همبستگی میان مقادیر مختلف این دیتاست را با استفاده از نمودار heatmap بدست آورده و در صورت وجود رابطه قوی میان دو متغیر غیرهدف، آن را گزارش کنید.
- ۷- نمودار مقادیر ستون battery\_power را به صورت هیستوگرام نمایش دهید.

سوال ۶ - از دیتاست Student که در اختیارتان قرار داده شده است برای حل سوالات زیر استفاده نمایید:

- ۱- نخست دیتاست را با استفاده از کتابخانه pandas خوانده و تبدیل به دیتافریم نمایید.
- ۲- پیش بینی می شود که math\_score دانشجویانی که پسر هستند از دانشجویانی که دختر هستند بیشتر باشد. درستی یا نادرستی این پیش بینی را با رسم نمودار هیستوگرام فیلد math\_score برای هر گروه و نیز محاسبه میانگین، مینیمم و ماکزیمم نمرات ریاضی هر گروه بررسی کرده و نشان دهید.
- ۳- با استفاده از نمودار Pie فراوانی مقادیر race/ethnicity را نشان دهید.
- ۴- فیلد جدیدی با عنوان total\_grade ایجاد کرده و میانگین نمرات writing\_score ، reading\_score و math\_score هر دانشجو را در آن ذخیره نمایید.
- ۵- جدول همسانی (contingency table) دو متغیر gender و parental level of education را بدست آورده و نمایش دهید.

سوال ۷ (اختیاری) - از دیتاست Housing که در اختیارتان قرار داده شده است برای حل سوالات زیر استفاده نمایید:

- ۱- نخست دیتاست را با استفاده از کتابخانه pandas خوانده و تبدیل به دیتافریم نمایید.
- ۲- سطرهای شامل مقادیر Null را در این دیتاست حذف کنید.
- ۳- برای ستون ocean\_proximity تمامی مقادیر یکتا را به همراه تعداد هر مقدار نمایش دهید.
- ۴- دو ویژگی longitude و latitude را با استفاده از داده های مکانی بر روی نقشه نمایش دهید و تراکم مناطقی که تعداد خانه بیشتری در آن ها وجود دارد یا نیز روی نقشه مشخص نمایید.

نحوه تحویل: سوالات تئوری را به صورت تایپ شده و در قالب یک فایل PDF تحویل دهید. به علاوه هر یک از سوالات عملی را در قالب یک فایل ipynb به همراه نتایج قرار داده و فایل را به صورت Qn نام گذاری نمایید که n شماره سوال مربوطه می باشد. در انتها فایل های پایتون را به همراه فایل PDF تماما در قالب یک فایل zip نامگذاری شده به صورت NAME\_STUDENTID در سامانه درس بارگذاری کنید. برای سوالات عملی توضیحات خود را به صورت Markdown در فایل پایتون بنویسید.

*The greatest glory in living lies not in never falling, but in rising every time we fall .*