به نام خدا

دانشگاه تهران

دانشکدگان فنی

دانشکده مهندسی برق و کامپیوتر

درس داده کاوی

تمرین ۱

اسفند ماه ۱۴۰۱

فهرست

سوال ۱ ..... ۳
سوال ۲ ..... ۴
سوال ۳ ..... ۵
سوال ۴ ..... ۶
ملاحظات (حتما مطالعه شود) ۷

سوال ۱

مجموعه داده های زیر را در نظر بگیرید:

- مجموعه داده‌ی دما و رطوبت
- مجموعه داده‌ی جنسیت (مرد یا زن) و قد (به سانتی‌متر)
- مجموعه داده‌ی سن (بر حسب سال) و اینکه آیا فرد دارای شرایط پزشکی است یا خیر
- مجموعه داده‌ی نوع خودرو (سدان، شاسی بلند یا کامیون)، اسب بخار و مصرف سوخت در صد کیلومتر

برای هر یک از مجموعه داده‌ها، ابتدا مشخص کنید خصیصه، گسسته، پیوسته یا باینری است و نوع آن را نیز مشخص نمایید. سپس مناسب ترین نوع نمودار را برای مصورسازی رابطه بین متغیرها انتخاب کنید. دلیل انتخاب خود را توضیح دهید و در مورد مزایا و محدودیت‌های انتخاب خود صحبت کنید.

سوال ۲

مجموعه داده‌ی زیر را در نظر بگیرید:

۱۲, ۱۹, ۸, ۲۳, ۱۵, ۱۷, ۱۰, ۵, ۱۴, ۷, ۱۸, ۲۰, ۱۱, ۱۳, ۱۶, ۹, ۲۲, ۲۵, ۲۴, ۶, ۲۱

Bining را روی این مجموعه داده با استفاده از هر دو روش عرض و عمق مساوی انجام دهید. برای عرض مساوی، از عرض پنج Bin و برای عمق مساوی، از سه Bin استفاده کنید. Binهای به دست آمده را برای هر روش مقایسه کنید و مزایا و محدودیت‌های هر روش را مورد بحث قرار دهید. علاوه بر این، یک Histogram برای مجموعه داده اصلی ایجاد کنید و Binهای حاصل را برای هر روش روی Histogram قرار دهید.

سوال ۳

مجموعه داده زیر را در نظر بگیرید:

ID	Age	Income	Savings
1	23	25000	10000
2	45	45000	20000
3	27	30000	15000
4	52	55000	30000
5	32	40000	18000
6	47	50000	25000
7	38	35000	12000
8	31	27000	8000
9	41	42000	22000
10	35	38000	16000

۱. برای هر ویژگی (سن، درآمد و پس‌انداز)، چارک اول (Q۱) و چارک سوم (Q۳) را محاسبه کنید.
۲. یک نمودار جعبه‌ای برای ویژگی "Savings" ایجاد کنید.
۳. یک نمودار پراکندگی ایجاد کنید که رابطه بین "سن" و "درآمد" را نشان می دهد.
۴. یک نمودار Q-Q برای ویژگی "درآمد" ایجاد کنید.
۵. ضرایب همبستگی را برای همه جفت ویژگی‌ها محاسبه و تحلیل کنید.

سوال ۴

مجموعه داده زیر شامل ۶ مشاهدات با سه ویژگی قد، وزن و سن را در نظر بگیرید:

ID	Height (cm)	Weight (kg)	Age (years)
1	165	70	30
2	170	65	28
3	155	45	35
4	180	90	40
5	160	50	25
6	175	75	32

۱. با استفاده از تکنیک نرمال‌سازی min-max، مجموعه داده را نرمال کنید.
۲. با استفاده از تکنیک decimal scaling، مجموعه داده را نرمال کنید.
۳. با استفاده از تکنیک z-score مجموعه داده را نرمال کنید.
۴. با استفاده از داده‌های نرمال شده min-max، مشابه‌ترین رکورد با رکورد پرس‌وجو (۱۷۵، ۸۰، ۳۵) را با استفاده از معیارهای فاصله اقلیدسی، منهتن و Supremum پیدا کنید.
۵. سوال ۴ را با استفاده از داده های نرمال شده با Decimal Scaling تکرار کنید.
۶. سوال ۴ را با استفاده از داده های نرمال شده با z-score تکرار کنید.

ملاحظات (حتما مطالعه شود)

تمامی نتایج شما باید در یک فایل فشرده با عنوان DM_HW1_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.
- مهلت تحویل تمرین به هیچ عنوان تمدید نخواهد شد. تمرین تا یک هفته بعد از مهلت تعیین شده با جریمه تحویل گرفته می‌شود که جریمه تاخیر تحویل تمرین تا یک هفته ۳۰درصد است.
- توجه کنید این تمرین باید به صورت تک نفره انجام شود و پاسخ‌های ارئه شده باید نتیجه فعالیت فرد نویسنده باشد (همفکری و به اتفاق هم نوشتن تمرین نیز ممنوع است). در صورت مشاهده تقلب به همه افراد مشارکت کننده، نمره تمرین صفر و به استاد نیز گزارش می‌گردد.
- در صورت بروز هرگونه مشکل با ایمیل زیر در ارتباط باشید:

mailto:Hoomanshirvani@ut.ac.ir

مهلت تحویل بدون جریمه: ۱۴۰۱/۱۲/۱۲

مهلت تحویل با تاخیر، با جریمه ۳۰ درصد: ۱۴۰۱/۱۲/۱۹