تمرین سری دوم – تحلیل آماری کاربردی

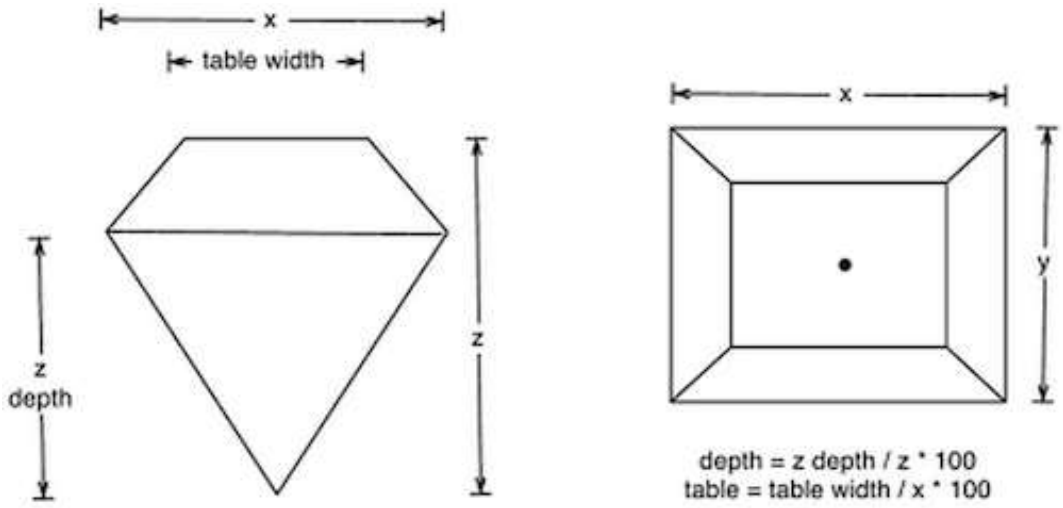
نکات مهم:

- برای تحویل تمارین گزارشی با استفاده از RMarkdown ایجاد نموده که در آن به صورت کامل صورت سوال، کد مربوطه، خروجی کد و توضیحات مربوطه آورده شود و سپس در سامانه آموزش مجازی بارگذاری شود. برای بارگذاری، ابتدا فایل word خروجی RMarkdown را تهیه کرده و سپس آن را pdf نموده و فایل pdf را به همراه فایل Rmd در سامانه بارگذاری کنید. صورت سوال را می توانید به صورت یک عکس در گزارش بیاورید. نام فایل ها حتما شامل شماره تمرین و شماره دانشجویی باشد (به عنوان مثال، HW2- 9913418023).
- دقت کنید که حتما برای پاسخ به تمامی قسمت های سوالات کدی نوشته و پاسخ را به صورت خروجی کد بیاورید.
- توصیه می شود که در تمرین سری دوم برای رسم نمودارهای مربوطه از کتابخانه ggplot2 و برای کار با داده ها از کتابخانه dplyr نیز استفاده شود.

اطلاعات در مورد مجموعه داده diamonds:

مجموعه داده diamonds که در این تمرین استفاده خواهیم کرد، شامل اطلاعات قیمت و کیفیت درباره تقریبا ۵۴۰۰۰ الماس است که در کتابخانه ggplot2 گنجانده شده است. برای دسترسی به این مجموعه داده باید ابتدا این کتابخانه بارگیری شود. (library(ggplot2))

این مجموعه داده حاوی اطلاعاتی در مورد قیمت الماس (در سال ۲۰۰۸ بر حسب دلار) و همچنین ویژگی های مختلف الماس است، که برخی از آنها بر قیمت آنها تأثیر می گذارند: قیراط (carat)، تراش (cut)، رنگ (color) و وضوح (clarity) و همچنین برخی از اندازه گیری های فیزیکی (عمق (depth)، جدول (table)، x، y، و z). شکل زیر نشان می دهد که این اندازه گیری ها چه چیزی را نشان می دهند.

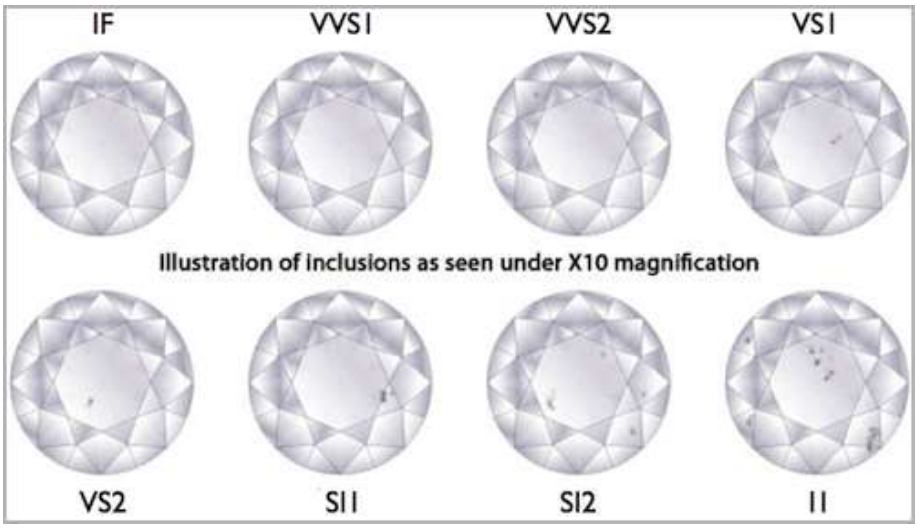


قیراط (carat) واحد جرمی برابر با ۲۰۰ میلی گرم است و برای اندازه گیری سنگ های قیمتی و مروارید استفاده می شود. درجه تراش (cut grade) معیاری عینی از عملکرد نور الماس یا چیزی است که ما عموماً آن را درخشش می‌دانیم.

تمرین سری دوم – تحلیل آماری کاربردی

شکل زیر درجه بندی رنگ الماس ها (color grading of diamonds) را نشان می دهد:

در نهایت، شکل زیر درجه بندی شفافیت الماس (clarity grading of diamonds) را نشان می دهد:



قسمت اول تمرین: EDA

۱. متغیر قیمت (price) چه نوع متغیری است؟ کمی یا کیفی؟ با استفاده از معیارهای عددی، تمایل مرکزی و پراکندگی آن را محاسبه و تحلیل کنید.

۲. یک هیستوگرام برای متغیر قیمت الماس تهیه کنید و با توجه به آن شکل توزیع این متغیر را تحلیل کنید. تعداد بازه های هیستوگرام را چند بار تغییر دهید و سعی کنید یک مقدار مناسب برای تعداد بازه ها پیدا کنید.

۳. متغیر رنگ (color) چه نوع متغیری است؟ آیا تمام رنگ های الماس (که در شکل بالا نشان داده شده است) در مجموعه داده وجود دارد؟ کدام رنگ بیشتر در مجموعه داده وجود دارد؟ با استفاده از یک جدول فراوانی، درصد هر یک از رنگ ها را در مجموعه داده مشخص کنید و با استفاده از یک نمودار میله ای آن را تشریح و تحلیل کنید.

۴. الماس های کدام درجه از تراش، بیشترین عمق (depth) را دارند؟ به طور متوسط، با افزایش یا کاهش درجه تراش، عمق افزایش یا کاهش می یابد؟

تمرین سری دوم – تحلیل آماری کاربردی

۵. توزیع متغیر قیمت را برای درجه های تراش مختلف مقایسه کنید. آیا چیزی غیرعادی به نظر می رسد؟ توضیح دهید.
۶. با استفاده از یک جدول توافقی و یک نمودار میله ای رابطه بین دو متغیر رنگ و درجه تراش را توصیف کنید.
۷. با استفاده از یک نمودار پراکندگی رابطه بین دو متغیر قیمت و قیراط را تشریح کنید. آیا الگویی در نمودار دیده می شود؟ سپس این نمودار را فقط برای الماس های دارای قیمت کوچکتر یا مساوی ۱۰۰۰ دلار، با و بدون استفاده از jittering رسم کنید. آیا استفاده از jittering سبب تغییر در شکل ظاهری نمودار و تحلیل آن می شود؟ توضیح دهید.
۸. رابطه بین دو متغیر قیمت و قیراط را بر اساس متغیر درجه تراش تشریح کنید. سعی کنید از ویژگی های ظاهری مختلف برای رسم نمودارها استفاده کنید.
۹. با استفاده از نمودار پراکندگی، رابطه بین دو متغیر قیمت و قیراط را بر اساس متغیر درجه شفافیت تشریح کنید. سپس، متغیرهای قیمت و قیراط را در مقیاس لگاریتمی در نمودار پراکندگی وارد کنید و نمودارها را تحلیل کنید.
۱۰. **(تمرین تشویقی)** سعی کنید ۲ مورد تحلیل دیگر با استفاده از ابزارهای مختلف EDA بر روی این مجموعه داده ارائه دهید. خلاق باشید و در هر یک از تحلیل ها حداقل ۳ متغیر را دخیل کنید.

قسمت دوم: آمار و احتمال (در تمامی قسمتها سطح معنادار بودن را ۵ درصد فرض کنید)

۱۱. آیا رابطه معناداری میان دو متغیر رنگ و درجه تراش وجود دارد؟
۱۲. آیا رابطه معناداری میان دو متغیر قیمت و درجه تراش وجود دارد؟ اگر بله، مشخص کنید که قیمت بین کدام سطوح متغیر درجه تراش متفاوت است؟

برای حل قسمت های بعدی، زیر مجموعه ای از داده ها را در نظر بگیرید که مقدار متغیر قیراط الماس ها کوچکتر از ۰.۵ باشد و برای این زیرمجموعه به قسمت های بعدی پاسخ دهید.

۱۳. آیا میانگین متغیر قیمت برای الماس های دارای درجه تراش Fair و Good اختلاف معناداری دارد؟
۱۴. آیا نسبت الماس هایی که دارای قیمت بیشتر از ۱۰۰۰ دلار هستند، برای درجه تراش های Fair و Good اختلاف معناداری دارد؟

(تمرین تشویقی) تابعی تحت عنوان z.test بنویسید که مشابه تابع t.test آزمون‌های فرض مربوط به میانگین یک و یا دو توزیع نرمال را با فرض معلوم بودن واریانسهای جامعه انجام دهد. دقت کنید که در خروجی تابع حتما p-value و فاصله اطمینان نیز محاسبه شود. سپس با استفاده از شبیه سازی، توزیع p-value در زمانی که فرض صفر برقرار است را برای آزمون فرض برابری میانگین دو توزیع نرمال با واریانس های معلوم حدس بزنید. برای این کار، در هر مرتبه شبیه سازی، ابتدا نمونه های تصادفی جدا از یک توزیع نرمال بگیرید و سپس با استفاده از تابع z.test مقدار p-value را محاسبه کنید و در نهایت با انجام تعداد زیادی شبیه سازی و رسم هیستوگرام مربوط به مقادیر p-value محاسبه شده، توزیع آن را حدس بزنید.

موفق باشید– شادمان