فاز اول پروژه

درس مبانی داده کاوی

نیمسال اول تحصیلی 1403

بخش اول: شناخت مجموعه داده

بااستفاده از دیتاست انیمیشن، ویژگی‌های عددی مانند vote_average، vote_count، runtime، revenueو budgetرا بررسی کرده و مقادیر زیر را برای هر کدام محاسبه کنید.

نام ویژگی	نوع	بازه مقادیر	Min	Max	Mean	Mode	Median	Outlier

برای ویژگی‌های عددی مشخص شده، نمودار Box Plot را رسم کنید و مقادیر پرت را شناسایی کنید.

مصورسازی داده‌ها :با استفاده از ابزارهای مصورسازی، نمودارهایی برای نمایش توزیع ویژگی‌های مهم ترسیم کنید. علاوه بر موارد زیر، 3 مورد دلخواه دیگر نیز با استفاده از نمودار تحلیل کنید.

- با استفاده از ستون‌های revenueو budget، رابطه میان بودجه و درآمد فیلم‌ها را تحلیل کنید و با رسم نمودارهایی مانند scatter plot بررسی کنید که آیا بودجه بیشتر منجر به درآمد بیشتر می‌شود یا خیر.
- ستون‌های vote_averageو vote_countرا با هم مقایسه کنید وبا استفاده از نمودارهایی مانند histogram بررسی کنید آیا فیلم‌هایی که تعداد آراء بیشتری دارند، میانگین امتیاز بالاتری کسب کرده‌اند.
- از ستون release_dateبرای تحلیل زمان‌بندی انتشار فیلم‌ها استفاده کنید با استفاده از نمودار خطی بررسی کنید در هر سال چه تعداد فیلم‌های انیمیشنی منتشر شده است.

موارد زیر را بررسی و تحلیل کنید.

- تحلیل ژانرها :از ستونgenres برای شناسایی ژانرهای پرطرفدار استفاده کنید. ژانرها را دسته‌بندی کرده و بررسی کنید کدام ژانرها(باتوجه به تعداد فیلم ها) بیشترین امتیاز رای‌دهی (vote_average)و بیشترین درآمد (revenue) را دارند تا مشخص شود که کدام ژانرها بیشتر مورد استقبال قرار گرفته‌اند.
- تحلیل تولیدکنندگان :بررسی کنید که کدام شرکت‌های تولید (production_companies)بیشترین تعداد فیلم با امتیاز بالا را تولید کرده‌اند و آیا این فیلم‌ها از نظر محبوبیت، امتیاز یا درآمد موفق بوده‌اند؟
- از ستون adultاستفاده کنید تا فیلم‌هایی که برای بزرگسالان ساخته شده‌اند را شناسایی کنید و تحلیل کنید که آیا این فیلم‌ها موفقیت بیشتری (از نظر امتیاز، درآمد و محبوبیت) نسبت به فیلم‌های غیر بزرگسال دارند یا خیر.
- با استفاده از ستون‌های revenueو budget، نسبت درآمد به بودجه را برای هر فیلم محاسبه کنید و بررسی کنید کدام فیلم‌ها از نظر مالی موفق‌تر بوده‌اند.
- با استفاده از ستون spoken_languages، فیلم‌ها را بر اساس زبان‌هایی که در آن‌ها صحبت می‌شود، دسته‌بندی کنید و بررسی کنید کدام زبان‌ها بیشتر مورد استفاده قرار گرفته‌اند.

بخش دوم: ارزیابی کیفیت داده

با بررسی مقادیر گم شده (missing) ، داده های پرت (outlier) ، ناهمسانیها و خطاهای موجود، کیفیت مجموعه داده را در ستون‌های مهمی مانند revenue، budgetو release_date ارزیابی کنید و برای این مرحله سعی کنید مورد ذیل را برای آن انجام دهید.

- بررسی کیفیت داده‌ها بر اساس مدل کیفیتISO 25012 باتوجه به فاکتورهای زیر:

نام ویژگی	تعداد رکورد	تعداد مقدار Null	Accuracy	Completeness	Validity	Currentness	Consistency

نکات تکمیلی:

- Consistency :برای بررسی سازگاری داده‌ها، می‌توان بررسی کرد که آیا مقادیر ستون‌هایی که به یکدیگر مرتبط هستند (مانند budget و revenue)با هم مطابقت دارند یا نه. به عنوان مثال، اگر budgetخالی باشد اما revenue مقدار داشته باشد، داده ناسازگار است. لطفا موارد دیگری را مانند مثال ذکر شده بررسی کنید.
- Currentness: برای ارزیابی تازگی داده‌ها، ستون release_date را در بازه های ده ساله بررسی کنید . داده‌های مربوط به فیلم‌های جدیدتر وزن بیشتری دریافت می‌کنند.
- Validity: بررسی اعتبار داده‌ها با ارزیابی مقادیری که از لحاظ منطقی معتبر هستند انجام می‌شود.برخی از فرمت ستونها در فایل dictionary داده شده است.
- Accuracy: به نسبت داده های معتبر و دارای مقدار به کل تعداد داده ها گفته میشود.

با توجه به موارد زیر در جدول، اشکالاتی که در دیتاست وجود دارد را مشخص کنید و به صورت مختصر درباره هر کدام توضیح دهید .

Single-Instance	Single-Schema

برای بهبود کیفیت داده مورد نظر، راهکارهای خود را ارائه نمایید و چه راه‌حل‌هایی برای پر کردن یا جایگزینی این مقادیر دارید.

بخش سوم: پیش‌پردازش داده‌ها(Preprocessing)

در این بخش شما باید داده هایی که در اختیار دارید را به فرمی ساختار یافته و تمیز تبدیل نمایید و در یک قالب مناسب برای تجزیه و تحلیل تبدیل کنید .

*در نظر داشته باشید پیش‌پردازش شما باید به گونه‌ای باشد که در فاز بعدی برای موارد مطرح‌شده زیر مناسب باشد:

- فیلم‌هایی که ویژگی‌های مشابهی از نظر محتوا دارند در یک خوشه قرار میگیرند :فیلم‌هایی که ژانرها، موضوعات یا محتوای مشابه دارند، باید در یک دسته قرار بگیرند.
- خوشه‌بندی فیلم‌ها بر اساس زمان انتشار :فیلم‌ها باید بر اساس زمان انتشار (به‌عنوان مثال، بر اساس فصل یا سال) خوشه‌بندی شوند.
- تحلیل شکست یا موفقیت فیلم‌ها بر اساس بودجه و محبوبیت :فیلم‌ها باید بر اساس میزان بودجه و محبوبیت آن‌ها تحلیل شوند تا بتوان شکست یا موفقیت فیلم‌ها را بر این اساس بررسی کرد.
- دقت کنید که طبقه بندی نهایی روی امتیاز فیلم هست.

موارد زیر برخی از اقداماتی است که در این بخش باید انجام دهید.

- پر کردن داده‌های گم‌شده :در ستون‌هایی که داده‌های گم‌شده(missing) دارند از روش‌هایی مانند یانگین، مد، میانه و یا رگرسیون استفاده کنید.
- نرمال‌سازی داده‌ها :ویژگی‌های عددی مانند vote_average، runtimeو popularityرا نرمال‌سازی کنید تا مناسب برای تحلیل شوند.
- شناسایی و حذف مقادیر پرت :مقادیر پرت(outlier) شناسایی شده در بخش اول را حذف یا تغییردهید.
- ایجاد ویژگی‌های جدید :با ترکیب ویژگی‌های موجود ویژگی‌های جدیدی را تعریف کنید.
- تبدیل داده‌ها :برخی داده‌های متنی را به داده‌های عددی تبدیل کنید .
- در صورت نیاز داده های عددی به داده های categorical تبدیل شوند .
- برای داده های متنی مثل ستون overview در صورت نیاز عملیات stemming, lemmetizing و حذف stopwords انجام شود. (برای این کار می توانید از کتابخانه nltk استفاه کنید)

نکات تحویل :

- پروژه در گروه های **حداکثر** دو نفری پیاده سازی شود.
- فایل ها باید در قالبstudentOneName-studentTwoName-Phase1.zip ارسال شود.

- ارسال فایل تنها از طریق سامانه **VU** مورد قبول بوده و فایل های ارسال شده در تلگرام و... تصحیح نخواهد شد.
- تمامی مراحل انجام پروژه و نتایج تحلیل‌ها و پیش‌پردازش‌های انجام شده را به صورت کامل در یک **گزارش** ارائه دهید.