پروژه‌ی پایتون – کلاس کاربرد کامپیوتر در مهندسی صنایع

کمپانی اسپاتیفای که یک سایت و اپلیکیشن استریم موزیک است، اطلاعات مربوط به آهنگ‌های برتر دهه‌ی دوم قرن ۲۱ خود را در طول زمان جمع آوری و ذخیره کرده که این اطلاعات در فایل Spotify.xlsx درخدمت شما قرار گرفته است. در بین این داده‌های جمع آوری شده، اطلاعات مشخصی از هر آهنگ آورده شده است. در ستون اول نام آهنگ، در ستون دوم خواننده‌ی موزیک و در ستون سوم سبک موسیقی آورده شده است. در ستون بعدی نیز سال انتشار هر قطعه ذکر شده است. در ادامه ۷ مشخصه از ویژگی‌های آهنگ آورده شده است که برخی از این ویژگی‌ها مانند beats per minute(bpm) یا dB و یا Duration به صورت عددی قابل اندازه گیری هستند. برخی نیز به صورت عددی قابل محسابه نیستند و توسط شاخص‌های کلیدی تعیین شده در شرکت و توسط کارشناسان موسیقی، عددی به آن‌ها نسبت داده می‌شود. (مانند Danceability)

در ستون آخر (Popularity) عددی به عملکرد آهنگ در سایت داده شده است. این شاخص که با توجه به میزان استریم‌ها و یا قرار گیری آهنگ در پلی‌لیست‌ها به موزیک‌ها نسبت داده شده است، محبوبیت هر آهنگ را نشان می‌دهد و با توجه به آن می‌توان گفت که یک موزیک، عملکرد موفقی داشته است یا خیر.

مشکلی که در این فایل وجود دارد این است که برخی از اطلاعات با استاندارد امروز کمپانی نوشته نشده اند. در ابتدا از شما خواسته می‌شود که اطلاعات را بر اساس استاندارد امروزی تنظیم کنید و سپس جداول خاصی را از داده‌های اصلی استخراج کنید که از آن‌ها برای تحلیل عملکرد موزیک‌ها و شاخص‌ها استفاده خواهد شد. سپس این جداول را به طوری که اطلاعاتی اضافی نداشته باشند، داخل یک فایل اکسل ذخیره کنید و این فایل را به عنوان گزارش داده‌ها ارائه کنید.

۱. در این بخش ابتدا می‌خواهیم سبک موسیقی‌ها را به نحوه‌ی استاندارد امروزی در بیاوریم. در گذشته شاخه‌های مختلف سبک‌های پاپ، راک، هیپ‌هاپ و... به صورت جداگانه نوشته می‌شدند که این کار باعث ایجاد دسته‌های زیاد و سلب امکان تحلیل مناسب می‌شد. حال برای دیتای داده شده از شما می‌خواهیم که سبک‌های قبلی را به شکل زیر به سبک‌های جدید تبدیل کنید. اگر در قسمتی از نام سبک 'pop' وجود داشت، به جای تمام آن نام، مقدار 'Pop' قرار دهد. به عنوان مثال 'canadian pop' به 'Pop' تغییر کند. به همین ترتیب هر سبک که داخل اسم آن 'hip hop' ، 'rock' و یا 'rap' داشت، به ترتیب به 'Hip Hop' ، 'Rock' و 'Rap' تبدیل شوند. برای باقی مقادیر، اگر نام ژانر، دو یا چند قسمتی بود (یک کرکتر فاصله بین دو کرکتر حرفی وجود داشت) آن را به 'Other' تغییر دهد و اگر نام یک قسمتی بود، حرف اول آن را بزرگ و باقی حروف را کوچک کند.

توجه کنید که مقادیر جانشین (ثانویه) باید دقیقا طبق الگو نوشته شوند (از نظر بزرگ و کوچک بودن حروف) و مقادیر اولیه نیز باید به هر صورتی که نوشته شده باشند، تشخیص داده شوند. (به عنوان مثال باید 'DanCE pOP' نیز به عنوان پاپ شناسایی شود و به 'Pop' تبدیل شود.

۲. در این قسمت که بخش نهایی پاکسازی داده است، می‌خواهیم موزیک‌هایی که خواننده‌ی میهمان دارند به روشی که امروزه در پایگاه‌های داده نوشته می‌شوند، در بیایند. در گذشته فیت‌ها به دو روش نشان داده می‌شدند. یا در قسمت خواننده به شکل 'XXX feat. YYY' نوشته می‌شدند. یا در نام آهنگ، به شکل 'ZZZ (feat. YYY)'. (که در این مثال‌ها 'YYY' نام خواننده‌ی میهمان، 'XXX' نام خواننده‌ی اصلی و 'ZZZ' نام موزیک است.) شیوه‌ی امروزی نشان دادن یک فیت به این شکل است که در قسمت خواننده، نام خواننده‌ها به شکل 'XXX + YYY' و نام موزیک بدون نام خواننده‌ی میهمان باشد. برای این کار باید در سطرهایی که در نام خواننده 'feat,' در صورت وجود به '+' تبدیل شود و اگر در نام موزیک به شکل 'ZZZ (feat. YYY)' نوشته شده بود، نام موزیک به 'ZZZ' تغییر پیدا کند و نام 'YYY' مانند قسمت قبل به نام خواننده‌ی اصلی در ستون Artist اضافه شود.

(در این قسمت اصلاح نام خواننده و آهنگ‌هایی که نام خواننده‌ی میهمان آن در قسمت نام موزیک (Title) آمده است امتیازی است و نمره‌ی اضافی دارد و تنها از شما خواسته شده است که این اصلاح را برای ردیف‌هایی انجام دهید که نام خواننده میهمان در ستون Artists آمده است.)

۳. حالا که داده‌ها به روز شده اند، در قسمت اول گزارش، یک جدول کوچک شامل اطلاعات کلی آماری این داده‌ها را (مقادیر عددی و غیر عددی) تهیه کنید و آن را در شیتی با نام 'Describe' ذخیره کنید.

۴. آهنگ‌هایی که bpm آن‌ها بیشتر از ۹/۱۰ (نُه دهم) مقدار بیشینه bpm است را جدا کنید. سپس بر اساس bpm به صورت **نزولی** سورت کنید و در انتها ستون‌های مربوط به نام آهنگ، خوانندگان، سبک و bpm و Popularity را داخل شیت بعدی به نام 'bpm' سیو کنید.

۵. در این قسمت یک جدول بسازید که برای هر سال میزان dB متوسط را محاسبه کرده و این داده‌ها را بر اساس dB به صورت نزولی مرتب کرده و داخل شیت بعدی به نام 'dB-mean' بریزد.
۶. در قسمت بعد می‌خواهیم ببینیم کدام خوانندگان پای ثابت موزیک‌های برتر سالیانه‌ی ما بوده اند. برای این کار تعداد تکرار نام هر خواننده را محاسبه کرده و جدولی شامل تعداد دفعات حضور هر خواننده در لیست و نام آن خواننده (۲ ستونی) طراحی کنید. خواننده‌هایی که بیشتر یا مساوی ۳ حضور داشته اند را جدا سازی کنید و بر اساس تعداد دفعات حضور، سورت کنید. جدول حاصله را در شیتی جدید به نام 'Best-Artists' سیو کنید.
۷. حالا می‌خواهیم ببینیم کدام سال تعداد موزیک هیت بیشتری داشته ایم. برای این کار جدولی بسازید که تنها شامل آهنگ‌هایی با Popularity بزرگتر یا مساوی ۸۰ باشند. سپس این مقادیر را در شیتی به نام 'Hit-Songs' ذخیره کنید. سپس در شیت بعدی جدولی بسازید که نشان دهد در هر سال چه تعداد آهنگ هیت داشته ایم. نایم این شیت را 'Hit-Songs-Count' بگذارید.
۸. می‌خواهیم ببینیم که میزان وابستگی بین محبوبیت یک آهنگ و Accousticness آن چه میزان است. برای این کار پارامترهای خط رگرسیون برازش شده را برای آن حساب کنید. (عرض از مبدا و شیب) فکر می‌کنید اگر این شاخص در یک آهنگ افزایش پیدا کند، محبوبیت آن افزایش پیدا می‌کند؟ بسیار کوتاه دلیل خود را ذکر کنید
۹. در این بخش شرکت می‌خواهد بداند کدام یک از شاخص‌ها، برای پیش‌بینی میزان محبوبیت آهنگ قابل اطمینان تر است. شرکت تاثیرات مستقیم (رگرسیون خطی) مثبت یا منفی را برای شاخص‌ها بررسی می‌کند. برای این کار باید بفهمیم فرض خطی بودن برای رابطه‌ی میان شاخص مد نظر و Popularity به چه میزان صحیح است. در ابتدا شاخص‌هایی که قدر مطلق برآورد شیب خط رگرسیون برازش شده‌ی آن‌ها کمتر از ۰.۱ است را کنار می‌گذاریم چرا که شیب خط نزدیک به صفر به این معنی است که دو متغیر با یکدیگر رابطه‌ی خطی ندارند. سپس باید ببینیم خطای پیش‌بینی برای هر خط باقی مانده به چه اندازه است. برای این کار ستون‌هایی تحت عنوان 'epsilon2 WWW' (به جای WWW نام شاخص را قرار دهید. مثلا ('epsilon2 Accousticness')) تعریف کنید و مربع خطای پیش‌بینی توسط خط برازش شده را داخل آن بریزید. (به بخش راهنمایی در انتهای سوال توجه شود) سپس مجموع مربعات خطا (SSE) را برای شاخص‌های مد نظر محاسبه کنید و بگویید کدام یک از شاخص‌ها برای پیش‌بینی مقدار Popularity مناسب تر اند.

راهنمایی: پیش‌بینی میزان محبوبیت یک آهنگ توسط خط رگرسیون به شکل مقابل است.

$$\hat{Y}_i = b_1 X_i + b_0$$

که در آن b_0 عرض از مبدا خط برازش شده بین مقادیر ستون شاخص مدنظر و b_1 شیب آن خط است. همچنین X مقدار شاخص، و $\hat{Y}$ پیش‌بینی میزان Popularity است.

حال خطای پیش‌بینی برابر است با مقدار انحرافی که مقدار پیش‌بینی شده با مقدار واقعی دارد. یعنی برای ردیف i ام داریم:

$$\varepsilon_i = Y_i - \hat{Y}_i$$

که در آن Y_i مقدار دقیق Popularity برای آن ردیف است.

به طور خلاصه برای تعریف ستون جدید 'epsilon2 WWW' برآورد شیب و عرض از مبدا خط رگرسیون شده بین ستون WWW و Popularity را محاسبه کنید (و به عنوان مثال در $b_{1,w}$ و $b_{0,w}$ بریزید) و سپس مقادیر ستون جدید را به صورت زیر تعریف کنید:

$$\left(Popularity - b_{1,w} WWW - b_{0,w}\right)^2$$

و سپس مقادیر هر ستون را با دستور ()sum. جمع کنید، و با مقدار متناظر دیگر ستون‌های اپسیلون مقایسه کنید.

استفاده از p-value برای آزمودن فرض صفر $b_1 = 0$ و تصمیم گیری بر مبنای آن، نمره‌ی اضافی خواهد داشت. هر چه مقدار p-value کوچکتر باشد، آن خط رگرسیون برای پیش‌بینی قابل اتکا تر است. می‌توانید با سرچ کردن در اینترنت، تابع و کتاب‌خانه‌ی مربوط به p-value این فرض را پیدا کنید.

نکات تحویل:

- این پروژه به صورت گروهی است.
- گروه‌هایی که این پروژه را به عنوان پروژه‌ی اصلی انتخاب کرده اند باید ۹ سوال را پاسخ دهند و گروه‌هایی که این پروژه را به عنوان پروژه‌ی فرعی انتخاب کرده اند، ملزم به پاسخگویی به ۸ سوال ابتدایی هستند.
- فایل تحویلی باید شامل سه فایل اکسل، اسکریپت پایتون (با فرمت py) و یک فایل pdf باشد.
- در فایل pdf باید نام و شماره‌ی دانشجویی اعضای گروه نوشته شده باشد. خطوط کد مربوط به هر بخش از پروژه آورده شود و توضیح مختصری در مورد قسمت ۸ و ۹ داده شود. لطفا فایل خود را ساده و بدون استفاده از تمپلیت طراحی کنید.
- یک فایل اکسل شامل ۸ شیت به عنوان گزارش باید ارسال شود. که به ترتیب نام شیت‌ها باید عبارت باشد از: Describe ، bpm ، db_mean ، Best_Artists ، Hit_Songs ، Hit_Songs_Count ، data. توضیحات مربوط به ۷ شیت اول در صورت سوالات داده شده است. در مورد شیت data گروه‌های اصلی، کل جدول به علاوه‌ی ستون‌های اپسیلون (بخش ۹) را وارد کنند و گروه‌های فرعی کل جدول پس از پاکسازی داده‌ها (بخش ۱ و ۲)
- حذف کردن indexها هنگام سیو کردن شیت جهت مرتب سازی فایل اکسل نمره‌ی مثبت خواهد داشت. توجه کنید که در برخی جدول‌ها مانند describe و groupby ها ستون index وجود ندارد و استفاده از دستور مربوطه منتج به حذف شدن بخشی از جدول شما می‌شود.
- فایل پایتون تحویل داده شده باید به نحوی باشد که در صورت اجرا، همان اکسل تحویل داده شده را تولید کند. بنابراین از دستکاری فایل اکسل تحویلی پس از ساخت آن توسط کد، خودداری نمایید.
- دستورات مورد نیاز برای سوالاتی که نمره‌ی مستقیم دارند در کلاس آموزش داده شده اند با این حال محدودیتی برای استفاده از هر گونه کتابخانه، کدهای خارج از مباحث کلاس و یا کمک گرفتن از اینترنت وجود ندارد.
- در صورت بروز هر گونه سوال، پرسش خود را مطرح کنید تا بتوانید به درستی و به آسانی به سوالات پاسخ دهید و نمره‌ی کامل را دریافت کنید.

در صفحه‌ی بعد دستوراتی که در هر سوال به کمک شما می‌آیند آورده شده اند. در سایت کوئرا نیز یک فیلم کوتاه جهت توضیح پروژه قرار خواهد گرفت.

1- Import pandas, readexcel , import re , *تعریف تابع* , .lower() , apply , find/in

2- in.find , split , join , apply , *تعریف تابع*

3- 4- 5- 6- 7- describe, sort_values *, groupby , mean , count*

8- import sklearn.linear_model , .values , reshape , fit

9- *تعریف تابع, تعریف ستون جدید,* sum