Home Work

Data Analysis

NUMBER	DEADLINE	TOPICS
		Final Project

۱. این پروژه **ارائه ی آنلاین و فردی** داره که در تاریخ های ۱۷، ۱۸ و ۱۹ مرداد انجام میشه. که بعدا زمان ارائه رو خودتون انتخاب میکنید که باید سر زمان انتخاب شده ارائه رو شروع کنید.

۲. هر ارائه باید حداکثر **۱۰ دقیقه** باشه. برای این ارائه یک فایل **PowerPoint** از فاز ۲ در کنار فایل **Power BI** آماده کنید و زمان ارائه ی شما باید بین این دو بخش تقسیم بشه. دقت کنید که **فایل powerpoint تنها جواب به سوالات مطرح شده نیست** و باید بتونید برداشت هاتون رو توضیح بدید. بعد از این زمان، **۵ دقیقه** به پرسش و پاسخ، دریافت فیدبک و دریافت نمره‌ی این دو فاز میگذره. فاز اول و آخر به صورت جداگانه تصحیح میشه و نمراتش اعلام میشه.

۳. مواردی که باید تا **ددلاین** در کارپوشه، تاپیک تمرین ارسال کنید. (بدیهیه که در صورت عدم رعایت هر کدوم از موارد نمره ی اون فاز **صفر** در نظر گرفته میشه و قابل تغییر نیست)

a. فایل **PDF** از فاز ۱ با نام گذاری به صورت رو به رو:
sql_firstname_lastname

این فایل شاملِ موارد زیر باید باشه

- صورت هر سوال
- اسکرین شات از query به همراه پاسخ خروجی

b. فایل **powerpoint** فاز ۲ با نام گذاری eda_firstname_lastname

c. فایل **powerbi** فاز ۳ با نام گذاری powerbi_firstname_lastname

d. **لینک** repository گیت‌هاب که کدها و ریپورت رو روی اون پوش کردید

۴. در صورت مشاهده ی تقلب نمره ی صفر برای دو طرف منظور میشه

تبریک میگیم :) به پروژه ی پایانی رسیدید!

بعد از پشت سر گذاشتن بوت کمپ دیتا آنالیز، شما به عنوان یک تحلیلگر داده توسط یک زنجیره فروشگاه موسیقی که مجموعه‌ای وسیع از آلبوم‌های موسیقی، آهنگ‌ها و تعاملات کاربران را مدیریت می‌کنه، استخدام شدید. مدیریت مجموعه قصد داره از داده‌های خود برای کسب بینش در مورد فروش، رفتار کاربران و عملکرد کلی کسب‌وکار استفاده کنه. وظیفه شما تحلیل مجموعه داده‌ی داده‌شده و ایجاد یک گزارش جامع و داشبورده که به مدیریت در گرفتن تصمیمات آگاهانه کمک کنه.

اهداف پروژه

درک پایگاه داده و نوشتن کوئری‌ها SQL: درک طرح پایگاه داده و نوشتن کوئری‌های SQL برای استخراج داده‌های مرتبط.

تحلیل داده‌های اکتشافی(EDA): انجام تحلیل داده‌های اکتشافی برای کشف الگوها، روندها و ناهنجاری‌ها در داده‌ها.

تحلیل آماری: انجام تحلیل‌های آماری برای پشتیبانی از یافته‌های خود.

بصری‌سازی داده‌ها و ایجاد داشبورد: استفاده از Power BI برای ایجاد یک داشبورد تعاملی که بینش‌ها و معیارهای کلیدی را بصری‌سازی کند.

گیت: پوش کردن تمام پروژه در یک ریپازیتوری

فاز ۱: پایگاه داده

قبل از هر چیزی نیازه که پایگاه داده رو بسازید. دیتابیس مورد نیاز شما در لینک گیتهاب زیر قرار داره:

Chinook Database

فرقی نمیکنه که از Oracle یا MySQL یا سایر RDBMS ها استفاده میکنید، این دیتابیس برای رنج وسیعی از RDBMSها طراحی شده که انتخاب اون با خود شماست.

پس از ساختن جداول، ER Diagram دیتابیس رو به روش هایی که پروژه ی PowerBI بیان شد بکشید.

جداول Fact و Dim رو مشخص کنید و بگید که به نظر شما این دیتابیس به کدام یک از اسکماهای star، snow flake و galaxy نزدیکتر است؟ چرا؟

حالا، با نوشتن کوئری های SQL به سوال‌های پاسخ بدید:

۱) ۱۰ آهنگ برتر که بیشترین درامد رو داشتن به همراه درامد ایجاد شده
۲) محبوب ترین ژانر، به ترتیب از نظر تعداد آهنگ‌های فروخته شده و کل درامد
۳) کاربرانی که تا حالا خرید نداشتند
۴) میانگین زمان آهنگ ها در در هر آلبوم
۵) کارمندی که بیشترین تعداد فروش را داشته
۶) کاربرانی که از بیش از یک ژانر خرید کردند
۷) سه آهنگ برتر از نظر درامد فروش برای هر ژانر
۸) تعداد آهنگ‌های فروخته شده به صورت تجمعی در هر سال به صورت جداگانه
۹) کاربرانی که مجموع خریدشان بالاتر از میانگین مجموع خرید تمام کاربران است

فاز ۲: تحلیل داده ها

توی فاز دوم قراره که با انجام مراحل مختلف، از داده ی خام به دانش برسیم و بتونیم برداشت هامون رو از وضعیت فعلی بیزنس گزارش کنیم.

۱) در جلسه‌ی هشتم کلاس منتورینگ، اتصال به دیتابیس و اجرای مستقیم کوئری‌ها از طریق پایتون رو یاد گرفتید. پس تمام جداول رو در dataframe های جداگانه لود کنید.
۲) با استفاده از دستورات مرتبط که در pandas میشناسید، یک بررسی اولیه برای داده ها انجام بدید

۳) متغیرهای کلیدی از نظر شما کدام متغیرها هستند؟ توزیع اون‌هارو با استفاده از plotهای مناسب بررسی کنید
۴) به نظر شما چه پلات‌هایی به شما در فهم بیشتر داده‌ها کمک میکنن؟ از اون‌ها استفاده کنید و برداشت‌های خودتون رو گزارش کنید
۵) با استفاده از ۳ روش مختلف، نرمال بودنِ متغیرهایِ کلیدیِ عددی رو بررسی کنید
۶) با روش‌هایی که تا اینجا یاد گرفتید، داده های پرت متغیرهای مرحله‌ی قبل رو پیدا کنید. به نظر شما باید حذف بشن؟ اگر بله، با دلیل داده های پرت رو حذف کنید
۷) با آزمون فرض مناسب به سوالات زیر پاسخ بدید (سوال ها جدا هستند و ارتباطی ندارند):
- با pandas ۳ ژانر محبوب رو پیدا کنید. تفاوت میانگین قیمت این دو توزیع رو بررسی کنید
- وجود استقلال بین طول آهنگ و قیمت آن را بررسی کنید
- تعداد آهنگ های خریداری شده توسط کاربران زن و مرد رو بررسی کنید
- استقلال ژانر آهنگ و نوع رسانه رو بررسی کنید
- طراحی **سه** سوال بعدی با شما :)

۸) به سوالات زیر در مورد فاصله اطمینان، با اعداد و پلات های مناسب پاسخ بدید (استفاده ی مناسب از pandas مهمه):
- میانگین طول آهنگ‌ها در ژانرهای مختلف یکسانه؟ فاصله اطمینان ۹۵ درصدی برای میانگین طول آهنگ‌ها در هر ژانر را محاسبه کنید
- میانگین فروش در کشورهای مختلف چقدره؟ فاصله اطمینان ۹۵ درصدی برای میانگین فروش در هر کشور را محاسبه کنید
- میانگین تعداد آهنگ‌های خریداری‌شده توسط هر کاربر چقدره؟ فاصله اطمینان ۹۵ درصدی برای میانگین تعداد آهنگ‌های خریداری‌شده توسط هر کاربر را محاسبه کنید

فاز ۳: داشبورد

تا اینجا داده‌هارو به خوبی شناختیم. حالا قراره که یک داشبورد مدیریتی بسازیم که به صورت real time به دیتابیس متصل باشه.

مانند پروژه ی پیاده سازی شده در power bi:

دو صفحه ی جداگانه برای Home و Overview در نظر بگیرید. **دو** صفحه ی دیگر هم طراحی کنید که موجودیت های مختلف رو به تصویر بکشن.

در صفحه ی Home، از فرمت مناسب برای تکست استفاده کنید، تمام متن هارو در یک text ساده ننویسید

در سایر صفحات از حداقل یک اسلایدر استفاده کنید. در مجموع این صفحات، ۶ کارت و ۳ KPI تعریف کنید

در هر صفحه حداقل از ۳ ویژوال مختلف (به جز موارد بیان شده) استفاده کنید

هر صفحه عنوان مناسب به شکل متنی در بالای صفحه داشته باشه

در مجموع این صفحات حداقل ۵ measure یا column تعریف کنید

از theme رنگی مناسب استفاده کنید.

در نهایت، زیاد سرچ کنید و داشبورد های مشابه ببینید تا ایده بگیرید

فاز ۴: گیت

حالا تمام پروژه رو روی گیتهابتون بذارید. تمام پروژه رو یک جا پوش نکنید. هر جایی که حس کردید به پایان یک بخش رسیدید پوش رو انجام بدید.

یک markdown مناسب برای پروژه بنویسید

تا زمان اعمال نتایج نهایی repository رو پابلیک نگه دارید