تمرین سری ۱ واحد درسی داده کاوی

جناب آقای دکتر فراهانی
دستیار آموزشی : علی شریفی

۲۰ اسفند ۱۴۰۰

پیشاپیش سال نو بر شما مبارک باد. با امید سالی سرشار از سلامتی و برکت و شادی.
یا مقلب القلوب و الابصار، یا مدبر اللیل و النهار، یا محول الحول و الاحوال، حول حالنا الی احسن الحال
ای تغییر دهنده دلها و دیده ها، ای مدبر شب و روز، ای گرداننده سال و حالت ها، بگردان حال ما را به نیکوترین حال

توجه کنید شما میتوانید بر روی کگل یا کولب و یا کامپیوتر های شخصی خود کار کنید .
به جای دانلود و آپلود دیتاست در گوگل درایو برای استفاده در کولب میتوانید به شیوه زیر عمل کنید .
چگونه از دیتاست های کگل در کولب استفاده کنیم ؟
ددلاین تمرین تا ۱۷ فروردین ۱۴۰۱ می باشد.
نحوه تحویل پاسخ تمرین ها در ریپازیتوری متعلق به درس می باشد.

۱ تمرین

۱.۱ دیتاست شماره ۱

در این تمرین از دیتاست شرکت airbnb استفاده میشود و در زمینه اجازه منازل در شهر نیویورک در ۲۰۱۹ می باشد. لینک دیتاست از شما خواسته میشود به تسک های زیر را انجام دهید.

۱.۱.۱ تسک های اصلی

۱. پاکسازی داده از قبیل بررسی داده های از غیرموجود و یافتن داده های پرت.

۲. ارایه اطلاعات کلی در حالت تجمیعی در خصوص اگهی از قبیل تعداد آگهی ها ، تعداد آگهی ها در هر منطقه جغرافیایی ، بررسی شاخص های کلی قیمت و ... . لازم است حتما در این بخش از بحث مصورسازی داده ها استفاده کنید و این اشکال ایجاد شده را تفسیر کنید.

۳. بررسی صاحبان آگهی و گزارشی از تعداد خانه های مرتبط با هر صاحب آگهی

۴. اگر تعداد کامنت های برای یک آگهی را بتوان شاخصی از تعداد مشتریان در نظر گرفت مطلوب است یافتن صاحبان آگهی که بیشترین مشتری را دارا می باشند و بررسی علت های آن.

۵. مطرح کردن ۵ آزمون فرض دلخواه در داده ها و پاسخ گویی و تفسیر آنها (حداقل از ۳ آزمون فرض متفاوت استفاده کنید.)

۶. تلاش در ساخت مدل برای پیش بینی پارامترهایی همانند قیمت و ارایه این مدل ها و تفسیر آنها.

۲.۱.۱ تسک های امتیازی

این تسک ها، فرای تسک های اصلی می باشد و پاسخ گویی به آنها دارای نمره امتیازی می باشد.

۱. اضافه کردن اطلاعات اضافی به دیتاست از قبیل اطلاعات جغرافیایی ایستگاه های مترو و موزه ها ، فرودگاه ، ایستگاه قطار مرکزی و ... و بررسی اثر گذاری این شاخص ها بر آگهی از قبیل قیمت و تعداد مشتری و میزان رضایتمندی مشتریان و ...

۲. بررسی نقش زن بودن یا مرد بودن صاحب آگهی در قیمت و تعداد مشتریان و میزان رضایتمندی مشتریان.

۲.۱ دیتاست شماره ۲

در این تمرین از دیتاست از آگهی های استخراج شده از یکی از بزرگترین پلتفرم های املاک کشور آلمان استفاده میشود. لینک دیتاست

۱.۲.۱ تسک های اصلی

۱. پاکسازی داده از قبیل بررسی داده های از غیرموجود و یافتن داده های پرت . بررسی موارد مشابهت [۱]

۲. ارایه اطلاعات تجمیعی از تعداد آگهی و پارامترهای مختلف آگهی از قبیل تعداد آگهی ها در مناطق مختلف جغرافیایی ، اطلاعات در خصوص تعداد انواع خانه ها ، بررسی قیمت در مناطق مختلف جغرافیایی و ... . لازم است حتما در این بخش از بحث مصورسازی داده ها استفاده کنید و این اشکال ایجاد شده را تفسیر کنید.

۳. تلاش جهت مدل سازی قیمت ها بر اساس پارامتر های مختلف آگهی.

۴. استفاده از بحث multiprocessing در بخش پاکسازی و پیش پردازش داده ها و و بررسی runtime فرایندها.

۵. استفاده از dask و pyspark در بخش پاکسازی و پیش پردازش داده ها و بررسی runtime فرایندها.

۲.۲.۱ تسک های امتیازی

۱. استفاده از مهندسی ویژگی در جهت بهبود مدل ها .

۲. استفاده از dask و pyspark در بخش مدلسازی داده ها و مقایسه با حالت عدم استفاده از آنها.

[1] duplicate

۲ نحوه ارسال

تمامی تمرین ها طبق نحوه بیان شده در کلاس های حل تمرین در داخل گیت هاب تحویل گرفته میشوند.
دانشجویان گرامی نهایتا تا ۱۸ فروردین ماه فرصت دارند تا کارهای خود را موفق در گیت هاب قرار دهند.
با توجه به فرصت سه هفته ای قبلی جهت تمرین در کار کردن با گیت و بررسی ارسال ها در گیت هاب و بیان موارد موفق و غیر موفق، ارسال غیرموفق طبق دستورالعمل گفته به منزله عدم ارسال تمرین در نظر گرفته میشود.