به نام خدا

دانشگاه تهران

دانشکدگان فنی

دانشکده مهندسی برق و کامپیوتر

درس داده‌کاوی

تمرین عملی ۱

اسفند ماه ۱۴۰۰

* فهرست

مقدمه و شرح دادگان

ویروس کرونا یا کووید-۱۹ یک نوع سندرم تنفسی حاد با عامل ویروسی از خانواده کرونا ویروس‌ها می‌باشد که همه‌ی کشورهای جهان را در مدت زمان کوتاهی درگیر کرده است. با توجه به شیوع و میزان مرگ و میر بالای این بیماری و از سوی دیگر، احتمال اوج مجدد کووید-۱۹ خصوصاً به دلیل نبود درمان اختصاصی، آشنایی و بررسی اطلاعات مربوط به ویروس کووید-۱۹ اهمیت زیادی دارد.

در مجموعه داده‌ای که در اختیار داریم، اطلاعات مربوط به مبتلایان، تست‌های جدید، مرگ بیماران و برخی ویژگی‌های مربوط به کشورها به تفکیک روز آورده شده است که توضیح دقیق هر کدام از ویژگی‌ها را از اینجا می‌توانید بررسی نمایید.

در این تمرین، شما می‌بایست به بررسی این مجموعه داده و پیش‌پردازش و مصورسازی آن بپردازید و براساس مصورسازی‌های انجام شده، تحلیل خود را بیان نمایید.

پیش‌پردازش

پیش‌پردازش، یکی از مهم‌ترین گام‌ها در پروژه‌های داده‌کاوی است. رویکردهای مختلفی در زمینه‌ی مدیریت داده‌های گم شده و تبدیل داده‌ها به فرمت‌های دیگر مورد استفاده قرار می‌گیرد و انتخاب دقیق این رویکردها تأثیر مستقیمی در کیفیت نتایج نهایی دارد؛ لذا همواره می‌بایست بهترین رویکرد را شناسایی و اعمال نمود.

۱. تعداد داده‌های گم شده در هر ویژگی را مشخص کنید. سپس، با ذکر دلیل، رویکرد مورد استفاده خود را برای پر کردن داده‌های گم شده در هر ستون مشخص کرده و اقدام به تکمیل داده‌های گم شده کنید.

۲. دیتافریم دیگری درست نمایید که در آن، تعداد کیس‌های جدید، تعداد واکسینه‌های جدید، تعداد فوتی‌ها و جمعیت برای هر کشور به صورت تجمیع شده محاسبه شده باشد. (محاسبه‌ی جمع داده‌ها از ابتدا تا آخرین تاریخ موجود در مجموعه داده‌ها برای هر کشور)

۳. ستون جدیدی با اسم تاریخ شمسی ایجاد کنید و برای ایجاد آن، تاریخ میلادی را به شمسی تبدیل نمایید.

۴. با توجه به تعداد بالای ویژگی‌ها، آیا می‌توان از تعداد ویژگی‌ها کاست؟ (از معیار correlation می‌توانید استفاده کنید.)

۵. دیتافریم جدیدی درست نمایید که در آن صرفاً اطلاعات مربوط به کشور ایران قرار داده شده باشد.

۶. در دیتافریم ایران، ستونی ایجاد نمایید که در آن، ماه به عنوان یک ویژگی مستقل در نظر گرفته شده است.

۷. دیتافریم جدیدی ایجاد نمایید که مجموعه‌داده ایران را بر اساس ماه در سال ۲۰۲۱ تجمیع کند.

نمایش دادگان

یکی از مواردی که در داده‌کاوی بسیار مورد استفاده قرار می‌گیرد، مصورسازی داده‌ها می‌باشد که به کمک آن می‌توان درکی از مجموعه داده‌ی مورد نظر به دست آورد و همچنین، تحلیل‌های کاملی بر اساس نمودارهای به دست آمده، ارائه نمود.

در این بخش از تمرین قصد داریم به مصورسازی داده‌های ذکر شده بپردازیم:

۱. کدام کشورها بهترین و کدام کشورها بدترین عملکرد در مهار ویروس کرونا را داشته‌اند؟ با یک نمودار مناسب این مساله را بررسی نمایید و برداشت خود را از نتایج ذکر نمایید. (منظور از عملکرد، تعداد فوتی نسبت به کل جمعیت است.)

۲. می‌خواهیم تاثیر واکسیناسیون بر تعداد فوتی‌ها را بررسی کنیم. برای این کار فرض کنید الزام است که اطلاعات ۵ کشور را بررسی کنیم. شما کدام کشورها را برای مقایسه انتخاب می‌کنید؟ با یک نمودار مناسب این مساله را بررسی نمایید و برداشت خود را از نتایج ذکر نمایید.

۳. قصد داریم سرعت واکسیناسیون در کشورهای مختلف را بررسی کنیم. برای این کار فرض کنید الزام است که اطلاعات ۵ کشور را ارزیابی کنیم. شما کدام کشورها را برای مقایسه انتخاب می‌کنید؟ با یک نمودار مناسب این مساله را بررسی نمایید و برداشت خود را از نتایج ذکر نمایید.

۴. روند سختگیری در حوزه‌ی کرونا در ایران را در طول زمان بررسی کنید، توجه نمایید برای پاسخ‌گویی به این سوال براساس تحلیل خود می‌توانید از ویژگی یا ویژگی‌های دلخواه استفاده نمایید، تحلیل خود را بیان نمایید.

۵. با استفاده از دیتافریم تجمیع شده‌ای که ایجاد کردید، برای ویژگی تعداد فوتی‌های هر کشور نمودار BoxPlot رسم کنید و کشورهای پرت را شناسایی کنید و رویکرد مناسبی برای آن‌ها اتخاذ نمایید. با توجه به مقدار میانه و میانگین، چولگی نمودار به کدام سمت می‌باشد؟

۶. تاثیر ویژگی‌های تراکم جمعیت، میانگین سنی، وجود امکانات بهداشتی، تعداد تخت بیمارستان‌ها و شاخص پیشرفت انسانی را بر تعداد فوتی‌ها و تعداد کیس‌های جدید با رسم نمودار مناسب بررسی کنید.

۷. رابطه بین وضعیت اقتصادی کشورها و تعداد افراد واکسینه شده را بررسی کنید و تحلیل خود را بیان نمایید.

۸. در سال ۲۰۲۱ توزیع تعداد مبتلایان به تفکیک ماه را بررسی نمایید و تحلیل خود را ذکر نمایید.

سؤالات امتیازی

۱. تعداد فوتی‌های سه ماه اخیر کشورهای مختلف را به نسبت جمعیت آن‌ها بر روی نقشه نمایش دهید.

۲. تعداد فوتی‌ها و تعداد واکسینه شده‌های ایران را به صورت هفتگی تجمیع نمایید. سپس، در یک نمودار مناسب رسم کنید؛ اگر داده‌های مربوط به هفته کامل نشده است، باید این هفته آخر را به صورت متمایزی نمایش دهید. سپس، نمودار را تحلیل کنید. (در این سوال باید کل داده‌های مربوط به ایران در نظر گرفته شود.)

ملاحظات (حتما مطالعه شود)

تمامی نتایج شما باید در یک فایل فشرده با عنوان DM_CA1_StudentID تحویل داده شود.

- این فایل فشرده، بایستی حاوی یک فایل با فرمت PDF (گزارش تایپ شده)، یک پوشه به نام Codes (شامل کدهای نوشته شده) و یک پوشه برای مجموعه داده‌های ایجاد شده باشد.
- خوانایی و دقت بررسی‌ها در گزارش نهایی از اهمیت ویژه‌ای برخوردار است. به تمرین‌هایی که به صورت کاغذی تحویل داده شوند یا به صورت عکس در سایت بارگذاری شوند، ترتیب اثری داده نخواهد شد.
- در رسم نمودارها به رنگ‌ها، فونت‌ها، اندازه و... توجه شود. درصورتی که از زبان فارسی و فونتی غیر از فونت حالت پیشفرض برای متن‌های به کار رفته در نمودارها استفاده کنید، نمره مثبت برایتان محسوب می‌شود.
- گزارش به صورت تایپ شده در قالبPDF شامل شرح آزمایش‌های انجام شده، پارامترهای آزمایش، نتایج و تحلیل‌ها باشد. دقت داشته باشید که در تمامی تمرین‌ها، نمره‌ی اصلی به تفسیر و تحلیل شما تعلق می‌گیرد.
- مهلت تحویل تمرین به هیچ عنوان تمدید نخواهد شد. تمرین تا چهار روز بعد از مهلت تعیین شده با جریمه تحویل گرفته می‌شود که جریمه تاخیر تحویل تمرین تا چهار روز ۳۰ درصد است.
- توجه کنید این تمرین باید به صورت تک نفره انجام شود و پاسخ‌های ارئه شده باید نتیجه فعالیت فرد نویسنده باشد (همفکری و به اتفاق هم نوشتن تمرین نیز ممنوع است). در صورت مشاهده تقلب به همه افراد مشارکت کننده، نمره تمرین صفر و به استاد نیز گزارش می‌گردد.
- در صورت بروز هرگونه مشکل، با آقای شیروانی از طریق ایمیل زیر در ارتباط باشید:
mailto:hoomanshirvani@ut.ac.ir

مهلت تحویل بدون جریمه: ۱۵ فروردین ۱۴۰۱

مهلت تحویل با تاخیر، با جریمه ۳۰ درصد: ۱۹ فروردین ۱۴۰۱