

بسمه تعالی

شماره ۱	پروژه درس مبانی داده‌کاوی:
۱۴۰۰/۸/۱۵	تاریخ تحویل پروژه:
پروژه پیش پردازش داده‌ها	عنوان پروژه:
پایتون (Python)، آر (R)، متلب (MATLAB)	زبان برنامه نویسی:

توضیحات پروژه:

در این پروژه دانشجویان باید بر روی فایل مجموعه داده‌ای (آمار جهانی مبتلایان، مرگ و میرها، واکسیناسیون کووید-۱۹) که در اختیارشان قرار داده شده است، عملیات پیش پردازش داده‌ها را به صورت زیر انجام دهند:

۱. ابتدا باید ردیف داده‌هایی را که در آن‌ها صفتهای زیر مقدار تهی دارند (صفر را تهی در نظر نگیرید) را تبدیل به مقادیر صفر نمایید.

total_vaccinations	people_vaccinated	people_fully_vaccinated
--------------------	-------------------	-------------------------

نکته: به عنوان مثال نمونه‌ی زیر غیر قابل قبول است و این ردیف داده‌ها باید در مجموعه داده به صفر تبدیل شوند.

نمونه‌ی غیر قابل قبول:

total_vaccinations	people_vaccinated	people_fully_vaccinated
54000	54000	

نمونه‌های قابل قبول:

total_vaccinations	people_vaccinated	people_fully_vaccinated
0	0	0
54000	54000	0

نکته: همچنین، باید صفتهای مردم که به صورت کامل واکسینه شده اند (people_fully_vaccinated) و کل واکسیناسیون انجام یافته (total_vaccinations)، پس از اولین مشاهده مقدار عددی در این خانه‌ها، در صورت مشاهده خانه یا خانه‌های خالی بعد از آن‌ها، مقدار مشاهده شده آن‌ها تا مشاهده مقادیرهای بعدی باید ثابت باشند. به عبارت دیگر نباید مقدار صفر به داده‌های بعد از اولین مقدار مشاهده شده داد، تا زمانی که مقدار جدیدی در این خانه یا خانه‌ها دیده شده و بروز شود. به عنوان مثال نمونه‌ی زیر را می‌توان مشاهده کرد.

در نمونه‌ی زیر، تغییر مقدار انجام یافته توسط برنامه برای یک تغییر با رنگ قرمز، برای دو تغییر با رنگ زرد و برای سه تغییر با رنگ سبز مشخص شده است.

total_vaccinations	people_vaccinated	people_fully_vaccinated
0	0	0
8200	8200	0
8200	0	0
54000	54000	0
120000	120000	0
240000	240000	0
240000	0	0
504502	448878	55624
504502	0	55624
547901	470341	77560

۲. پس از انجام مرحله ۱، سه صفت (ستون) با نام‌های زیر را باید به مجموعه داده اضافه کنید. مقادیر خواسته شده با استفاده از صفت جمعیت (population) مربوط به هر کشور به صورت درصد برای هر یک از صفتهای جدید باید برای تمامی نمونه‌ها محاسبه و اضافه نمایید.

total_cases_per_population	total_deaths_per_population	people_fully_vaccinated_per_population	صفت
(total_cases * 100) / population	(total_deaths * 100) / population	(people_fully_vaccinated * 100) / population	فرمول

نکته: به عنوان مثال برای یک کشور با داده زیر را در نظر بگیرید. مقادیر این سه صفت برای این کشور در دو نمونه داده به صورت زیر بدست می‌آید.

فرض می‌کنیم جمعیت کشور الف ۲۵۰ نفر است. موارد کلی ابتلا، فوت و افرادی که به صورت کامل واکسینه شده‌اند برای این کشور مطابق جدول صفحه‌ی بعدی می‌باشد.

total_cases	total_deaths	people_fully_vaccinated
10	2	0
15	3	10
25	3	50

با انجام محاسبات، مقدار صفت‌های ذکر شده برای کشور الف مطابق جدول زیر بدست آمده است.

total_cases_per_population	total_deaths_per_population	people_fully_vaccinated_per_population
4 %	0.8 %	0 %
6 %	1.2 %	4 %
10 %	1.2 %	20 %

۳. پس از انجام مرحله ۲، با استفاده از آزمون آماری نمونه جفت شده تی-تست (Paired Sample T-Test)،

مشخص کنید آیا بین جفت دسته صفت‌های زیر تفاوت آماری معناداری وجود دارد یا خیر؟ در هر حالت مقادیر

بدست آمده را برای جفت دسته صفت‌های زیر نشان بدهید.

جفت‌ها
people_fully_vaccinated_per_population , total_cases_per_population
people_fully_vaccinated_per_population , total_deaths_per_population
people_fully_vaccinated , total_cases
people_fully_vaccinated , total_deaths

نکته: تفاوت آماری معنادار زمانی رخ می‌دهد که مقدار p-value بدست آمده از تی-تست بین جفت دسته صفت‌ها حداقل کمتر از مقدار

۰,۰۵ باشد.

۴. پس از انجام مرحله ۲، نمودارهای مقدار رشد صفت‌های مشخص شده زیر را بر اساس ترتیب صفت تاریخ

(date) موجود در دیتاست با یکدیگر مقایسه و با رنگ‌های متفاوت نمایش دهید.

صفت‌ها
total_deaths_per_population , people_fully_vaccinated_per_population , total_cases_per_population
total_deaths , people_fully_vaccinated , total_cases
total_vaccinations , people_vaccinated , people_fully_vaccinated
total_vaccinations , people_vaccinated , people_fully_vaccinated :Continent بر اساس
total_deaths , people_fully_vaccinated , total_cases :Continent بر اساس

۵. با استفاده از نمودار نقطه‌ای (Scatter Plot)، رابطه‌ی داده‌ها را بین صفت‌های مشخص شده زیر به صورت دو

به دو نمایش دهید.

صفت‌ها
people_fully_vaccinated و total_cases
people_fully_vaccinated و total_deaths
total_vaccinations و total_cases
total_vaccinations و total_deaths
total_vaccinations و people_fully_vaccinated
total_deaths و total_cases

توجه: مجموعه داده پردازش شده را به صورت کامل باید در قالب یک فایل با فرمت اکسل^۱ (یا مقادیر جداشده با ویرگول^۲) ذخیره نمایید. همچنین تمامی اشکال را باید ذخیره کنید. در زمان تحویل پروژه نیز، برنامه‌ای که نوشته‌اید را باید اجرا کرده تا کارهای ذکر شده را انجام دهد:

۱. پردازش هر مرحله را به ترتیب انجام داده و ذخیره کند. ۲. باید شکل‌ها را در هر مرحله به ترتیب ایجاد، نمایش و ذخیره کند.

موفق باشید.

^۱ .xlsx

^۲ .csv