به نام خدا

دانشگاه تهران

دانشکدگان فنی

دانشکده مهندسی برق و کامپیوتر

درس داده کاوی

تمرین کامپیوتری اول

نام و نام خانوادگی : حسین سیفی

شماره دانشجویی : ۸۱۰۱۰۰۳۸۶

فروردین ۱۴۰۱

پیش پردازش

سوال ۱

تعداد سطرهای خالی برای هر یک از ستون‌های مجموعه داده به شرح زیر می‌باشد:

Column	Misses	Column	Misses
iso_code	0	total_vaccinations	120658
continent	9917	people_vaccinated	122844
location	0	people_fully_vaccinated	125608
date	0	total_boosters	148296
total_cases	3030	new_vaccinations	128384
new_cases	3172	new_vaccinations_smoothed	81524
new_cases_smoothed	5156	total_vaccinations_per_hundred	120658
total_deaths	20843	people_vaccinated_per_hundred	122844
new_deaths	20803	people_fully_vaccinated_per_hundred	125608
new_deaths_smoothed	22902	total_boosters_per_hundred	148296
total_cases_per_million	3785	new_vaccinations_smoothed_per_million	81524
new_cases_per_million	3927	new_people_vaccinated_smoothed	82815
new_cases_smoothed_per_million	5905	new_people_vaccinated_smoothed_per_hundred	82815
total_deaths_per_million	21585	stringency_index	35774
new_deaths_per_million	21545	population	1072
new_deaths_smoothed_per_million	23638	population_density	18323
reproduction_rate	40569	median_age	28378
icu_patients	142246	aged_65_older	29866
icu_patients_per_million	142246	aged_70_older	29114
hosp_patients	141072	gdp_per_capita	27708
hosp_patients_per_million	141072	extreme_poverty	74799
weekly_icu_admissions	160232	cardiovasc_death_rate	29428
weekly_icu_admissions_per_million	160232	diabetes_prevalence	22287
weekly_hosp_admissions	154759	female_smokers	60027
weekly_hosp_admissions_per_million	154759	male_smokers	61476
new_tests	98630	handwashing_facilities	97352
total_tests	96692	hospital_beds_per_thousand	42485
total_tests_per_thousand	96692	life_expectancy	11016
new_tests_per_thousand	98630	human_development_index	29953
new_tests_smoothed	81978	excess_mortality_cumulative_absolute	159940
new_tests_smoothed_per_thousand	81978	excess_mortality_cumulative	159940
positive_rate	87046	excess_mortality	159940
tests_per_case	87609	excess_mortality_cumulative_per_million	159940
tests_units	79655		

همانطور که دیده شد در این سوال با مجموعه داده‌ای با فضاهای خالی بسیار زیاد روبرو هستیم و هدف ما پر کردن این فضاهای خالی به مناسب ترین روش است تا در نمودارهای رسم شده برای این مجموعه داده شاهد کمترین خطا و ناهنجاری باشیم. به طور کلی برای حل این مشکل با توجه به نوع داده‌ها و ماهیت هر ستون، استراتژی‌های متفاوتی را می‌توان در نظر گرفت و اعمال کرد. این استراتژی‌ها در ادامه معرفی می‌شوند و ستون‌هایی که از هرکدام از استراتژی‌ها استفاده کرده‌اند مشخص می‌شوند.

۱. تکنیک Linear Interpolation: برای ستون‌هایی که توالی دارند و به نظر می‌رسد که سطرهای خالی نشان دهنده جا ماندن یا از دسترس خارج شدن داده‌های هستند . در ستون‌هایی که از این روش برای پر کردن فضاهای خالی آن‌ها استفاده شده است یک بار

هم تکنیک "سطر بعدی" اعمال شده است تا از خالی ماندن خانه‌های ابتدای مجموعه داده به دلیل خاصیت تکنیک Linear InterPolation جلوگیری شود.

۲. سطر بعدی: برای بعضی ستون‌های جدول که در آن‌ها توالی مقادیر نقش مهمی دارد فرض جدیدی را بدین شکل در نظر گرفتیم که خالی بودن یه سطر از یک ستون به معنی این است که داده جدید وارد نشده است و در نتیجه باید از مقدار سطر قبلی استفاده کرد اما استفاده از سطر قبل محدودیت‌ها و مشکلاتی را به همراه دارد. مشکل اول این است که برای سطرهایی از ابتدای جدول که خالی هستند هیچ مقدار قبلی وجود ندارد و دومین مشکل وجود کشورهای متفاوت در مجموعه داده است. مشکل دوم بیان می‌کند در صورتی که سطر اول از کشور جدید خالی باشد، مقدار آخرین سطر کشور قبلی جایگزین جای خالی می‌شود که طبیعتا مقدار درستی نیست. بنابراین با منطقی مشابه و برای جلوگیری از مشکلات ذکر شده، از مقدار سطر بعدی برای پر کردن فضای خالی در برخی ستون‌ها استفاده می‌کنیم.

۳. میانگین: از این تکنیک برای ستون‌هایی استفاده می‌شود که مقادیر آن‌ها به ازای یک کشور خاص در هفته‌های متفاوت، یکسان هستند.

۴. دستی: برای برخی ستون‌ها مانند جمعیت، کشورهایی که ستون جمعیت خالی دارند را بررسی کردیم و مقادیر مناسب را به صورت دستی در جای خود قرار دادیم.

۵. حذف: برخی از ستون‌های این مجموعه داده دارای تعداد سطرهای خالی بسیار زیادی هستند و بیشتر از ۹۰ درصد داده‌های ممکن را از دست داده‌اند اما بنابر خواسته سوال در این بخش هیچ ستونی را حذف نمی‌کنیم.

در جدول زیر نام هر ستون و شماره تکنیک مورد استفاده برای پر کردن سطرهای خالی در آن‌ها مشخص شده است:

Column	**Technique**	**Column**	**Technique**
iso_code	-	total_vaccinations	2
continent	4	people_vaccinated	2
location	-	people_fully_vaccinated	2
date	-	total_boosters	2
total_cases	2	new_vaccinations	2
new_cases	1	new_vaccinations_smoothed	2
new_cases_smoothed	1	total_vaccinations_per_hundred	2
total_deaths	2	people_vaccinated_per_hundred	2
new_deaths	1	people_fully_vaccinated_per_hundred	2
new_deaths_smoothed	1	total_boosters_per_hundred	2
total_cases_per_million	2	new_vaccinations_smoothed_per_million	2
new_cases_per_million	1	new_people_vaccinated_smoothed	2
new_cases_smoothed_per_million	1	new_people_vaccinated_smoothed_per_hundred	2
total_deaths_per_million	2	stringency_index	3
new_deaths_per_million	1	population	4
new_deaths_smoothed_per_million	1	population_density	3
reproduction_rate	3	median_age	3
icu_patients	1	aged_65_older	3
icu_patients_per_million	1	aged_70_older	3
hosp_patients	1	gdp_per_capita	3
hosp_patients_per_million	1	extreme_poverty	3
weekly_icu_admissions	1	cardiovasc_death_rate	3
weekly_icu_admissions_per_million	1	diabetes_prevalence	3
weekly_hosp_admissions	1	female_smokers	3
weekly_hosp_admissions_per_million	1	male_smokers	3
new_tests	1	handwashing_facilities	3
total_tests	2	hospital_beds_per_thousand	3
total_tests_per_thousand	2	life_expectancy	1

new_tests_per_thousand	1	human_development_index	3
new_tests_smoothed	1	excess_mortality_cumulative_absolute	1
new_tests_smoothed_per_thousand	1	excess_mortality_cumulative	1
positive_rate	1	excess_mortality	1
tests_per_case	1	excess_mortality_cumulative_per_million	1
tests_units	1		

سوال ۲

به کمک قطعه کد زیر دیتافریم دیگری ایجاد می‌کنیم که شامل مقدار تجمیع شده تعداد کیس جدید، تعداد واکسینه جدید، تعداد فوتی‌ها و جمعیت برای هر کشور باشد.

```
countries = list(set(covid['location']))

new_cases = dict()

new_vaccination = dict()

new_deaths = dict()

population = dict()

for c in countries:

    new_cases[c] = 0

    new_vaccination[c] = 0

    new_deaths[c] = 0

    population[c] = 0

for i in tqdm(range(len(covid))):

    new_cases[covid.iloc[i]['location']] = new_cases[covid.iloc[i]['location']] + covid.iloc[i]['new_cases']

    new_vaccination[covid.iloc[i]['location']] = new_vaccination[covid.iloc[i]['location']] +
covid.iloc[i]['new_vaccinations']

    new_deaths[covid.iloc[i]['location']] = new_deaths[covid.iloc[i]['location']] + covid.iloc[i]['new_deaths']

    population[covid.iloc[i]['location']] = int(covid.iloc[i]['population'])
```

این قطعه کد بدین صورت عمل می‌کند که برای هرکدام از ویژگی‌های مطرح شده یک دیکشنری ایجاد می‌کند و کلیدهای این دیکشنری نام کشورها یا همان ستون Location در مجموعه داده هستند و مقدار متناظر با هر کلید مقدار ویژگی محاسبه شده برای آن کشور می‌باشد. سپس در قطعه کدی دیگر(که ذکر آن بی فایده است) این دیکشنری‌ها با یکدیگر تجمیع شدند تا یک دیتافریم واحد با ستون‌های خواسته شده را بسازند.

سوال ۳

با استفاده از تابع from_gregorian کتابخانه jdatetime مقدار تاریخ شمسی برای هر یک از سطرهای ستون date را محاسبه می‌کنیم و به مجموعه داده اضافه می‌کنیم.

سوال ۴

بله امکان حذف تعدادی از ستون‌ها وجود دارد چرا که تعدادی زیادی از ستون‌ها طبق تعریف پایگاه‌های داده مشتق شده محسوب می‌شوند(مانند New_cases_per_million که از ستون به مراتب مهم‌تر New_cases مشتق شده است.)

همچنین تعداد زیادی از ستون‌ها هم مقدار داده‌های از دست رفته زیادی(بیش از ۷۰ درصد داده‌ها) دارند و حتی با دقیق‌ترین روش‌های موجود نیز امکان پر کردن این فضاهای خالی با دقت قابل قبولی که نمودارهای دقیقی را حاصل شود ندارند.

سوال ۵

دیتافریم ایران را با بررسی ویژگی Location سطر به سطر مجموعه داده و انتخاب سطرهایی با مقدار ایران Iran ایجاد می‌کنیم.

سوال ۶

ویژگی ماه برای مجموعه داده ایران با شکستن ویژگی تاریخ شمسی از روی نمادهای "-" و انتخاب بخش دوم آن ایجاد می‌کنیم.

سوال ۷

برای تجمیع مجموعه داده ایران نسبت به ماه نیز همانند پر کردن فضاهای خالی برای هر ستون باید رویکرد مناسبی را اتخاذ کنیم اما خوشبختانه این کار را قبلا در سوال ۱ بخش پیش پردازش انجام داده‌ایم!

۱. ستون‌هایی که از روش Linear Interpolation و سطر بعدی استفاده می‌کنند را نسبت به ماه به دلیل خواص آن‌ها با یکدیگر جمع می‌کنیم.
۲. ستون‌هایی که از روش میانگین استفاده کردیم نسبت به ماه میانگین می‌گیریم.
۳. ستون‌های نام کشور، قاره، کد ایزو و جمعیت را مقداری تکراری و یکتا در تمام ماه‌ها استفاده می‌کنیم.