ياسمه تعالى

انتخاب، ساده سازی و متناسب سازی داده ها

۱-انتخاب چند کشور که داده های کاملی در مجموع دادگان دارند (حداقل ۵ کشور) ۲-فیلتر و ساده سازی داده های موجود متناسب هدف پژوهشی ۳- تطبیق مجموعه دادگان با یکدیگر

برای انتخاب ۵ کشور که دارای بیشترین اطلاعات ثبتی و زمانی باشند، ابتدا دیتاست اول (کودها) را فیلتر و مورد تجزیه و تحلیل قرار داریم. به عنوان مثال برای کشور افغانستان به اطلاعات زیر دست پیدا کردیم :



با همین منطق، کل کشورها را فیلتر می کنیم تا بعد از آن به انتخاب ۵ کشور حداکثری اقدام کنیم :

```
1- Afghanistan 322 All
2- Albania 554 All
3- Algeria 572 All
4- Angola 761 All
4- Cote 889
3- Antigua and Barbuda 387 All
4- Cote 1862 Al
```

```
161- United Arab Emirates 1406 All
162- United Kingdom 1388 All
163- United Republic of Tanzania 1148 All
164- United States of America 1592 All
165- Uruguay 1186 All
166- Uzbekistan 919 All
167- Venezuela (Bolivarian Republic of) 988 All
168- Viet Nam 1198 All
169- Yemen 562 All
170- Zambia 1657 All
171- Zimbabwe 937 All
```

بعد از مشخص کردن تعداد رکوردهای ثبت شده برای هر کشور و در نظر گرفتن رزولوشن زمانی، کشورها را سورت کردیم:

```
24- Brazil 1779 All
142- Slovenia 1626 All
48- Denmark 1621 All
164- United States of America 1592 All
149- Sweden 1575 All
149- Sweden 1575 All
158- Turkey 1566 All
124- Poland 1520 All
144- Spain 1516 All
73- India 1514 All
46- Czechia 1507 All
90- Lithuania 1483 All
79- Italy 1478 All
166- Greece 1477 All
150- Switzerland 1472 All
59- France 1466 All
190- Mexico 1462 All
```

اکنون کشورهای سورت شده را با دیتاست دوم (میزان تولید محصولات زراعی) مطابقت دادیم تا ببینیم برای کدام یک از این کشورها، در مجموعه دوم اطلاعات داریم که به لیست زیر رسیدیم :

```
24- Brazil 1779 All + 444
142- Slovenia 1626 All + No Data
48- Denmark 1621 All + No Data
164- United States of America 1592 All + 444
149- Sweden 1575 All + No Data
158- Turkey 1566 All + 444
124- Poland 1520 All + No Data
144- Spain 1516 All + No Data
143- Spain 1516 All + No Data
144- Grechia 1507 All + No Data
90- Lithuania 1483 All + No Data
90- Lithuania 1483 All + No Data
66- Greece 1477 All + No Data
66- Greece 1477 All + No Data
150- Switzerland 1472 All + No Data
43- Croatia 1472 All + No Data
49- France 1466 All + No Data
190- Mexico 1462 All + No Data
```

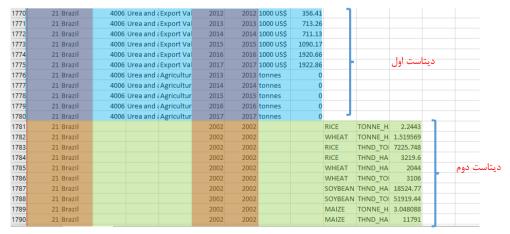
همان طور که در تصویر بالا پیداست برای بعضی کشورها در مجموعه دوم داده ای نداشتیم و برای برخی از کشورها هم ۴۴۴ دیتا داشتیم. و نهایتا با در نظر گرفتن همه این موارد ۵ کشور زیر را انتخاب کردیم:

```
24- Brazil 1779 All + 444 (BRA)
164- United States of America 1592 All + 444 (USA)
158- Turkey 1566 All + 444 (TUR)
73- India 1514 All + 444 (IND)
100- Mexico 1462 All + 444 (MEX)
```

در گام بعد ستون های غیر ضرور را از دیتاست اول حذف کردیم تا داده ها ساده تر شوند:

					×						×
1	Area Code	Area	Item Code	Item	Element Code	Element	Year Code	Year	Unit	Value	Flag

در گام آخر برای تطبیق دیتاست ها و سینک کردن آن ها به هم از اشتراک آن ها استفاده کردیم (سال ثبت و نام کشور)



ستون مشترک دیتاست ها ها

ستون مشترک دیتاست ها

* فایل نهایی به نام FinalDataset ارسال گردید.