

## گزارش پاکسازی داده ها

احمد رضا رستماني

سه فایل از دیتای تقسیم بندی شده شامل داده های مقدار و ارزش کود های وارد شده و میزان تولید کود در کشورهای آسیایی جهت پاکسازی بر عهده بنده بود که خروجی و دیتای پردازش شده در فولدر دیتا قرار گرفت.

بدلیل عدم وجود کشوری که در تمام سال های مورد بررسی مقدار کود خاصی برای آن 0 باشد و یا شامل null value بوده باشد تعداد ردیف ها یا در واقع مقادیر ثبت شده در قبل و بعد از پاکسازی بدون تغییر و برابر 11480 داده برای ارزش کودهای وارداتی، 11480 داده برای مقدار کودهای وارداتی و 2116 دیتای ثبت شده در تولید کود داشتیم. که از این مقدار به ترتیب 142 داده از ارزش واردات برابر 0 دلار، 11 داده ثبت شده برای مقدار کود وارداتی برابر 0 تن و 613 داده از قسمت تولید کود دارای مقدار 0 تن بودند.

قبل از پردازش، هر سه فایل دیتا شامل 12 ستون ( 6 ستون عددی و 6 ستون رشته ای ) بودند که از میان آنها ' Area ' 'Code', 'Item Code', 'Element Code', 'Element', 'Year Code', 'Unit', 'Flag', 'continent' بدلیل اینکه کاربردی در مسئله ما نداشتند حذف شده و 4 ستون Area, Value, Year, Item باقی ماندند که در مرحله بعدی نیز بدلیل نیاز به تبدیل این دیتا به دیتای سری زمانی ستون year بعنوان اندیس دیتا تنظیم شد.

نمونه قسمتی از فایل ورودی و خروجی:

47 AsiaIV.head()

47

	Area Code	Area	Item Code	Item	Element Code	Element	Year Code	Year	Unit	Value	Flag	continer
0	2	Afghanistan	4007	Ammonia, anhydrous	5622	Import Value	2011	2011	1000 US\$	2.69	R	Asia
1	2	Afghanistan	4007	Ammonia, anhydrous	5622	Import Value	2014	2014	1000 US\$	1.07	R	Asia
2	2	Afghanistan	4007	Ammonia, anhydrous	5622	Import Value	2017	2017	1000 US\$	1.24	R	Asia
3	2	Afghanistan	4003	Ammonium nitrate (AN)	5622	Import Value	2002	2002	1000 US\$	123.24	R	Asia
4	2	Afghanistan	4003	Ammonium nitrate (AN)	5622	Import Value	2007	2007	1000 US\$	7.18	R	Asia

41 AsiaIV.describe()

	Area Code	Item Code	Element Code	Year Code	Year	Value
count	11480.000000	11480.000000	11480.0	11480.000000	11480.000000	1.148000e+04
mean	133.740679	4014.393641	5622.0	2009.717073	2009.717073	3.102017e+04
std	70.527040	8.977773	0.0	4.570050	4.570050	1.652045e+05
min	1.000000	4001.000000	5622.0	2002.000000	2002.000000	0.000000e+00
25%	101.000000	4006.000000	5622.0	2006.000000	2006.000000	5.258750e+01
50%	121.000000	4014.000000	5622.0	2010.000000	2010.000000	6.571850e+02
75%	200.000000	4022.000000	5622.0	2014.000000	2014.000000	6.525047e+03
max	249.000000	4030.000000	5622.0	2017.000000	2017.000000	6.294714e+06

50 AsiaIV.head()

	Area	Item	Value
Year			
2011	Afghanistan	Ammonia, anhydrous	2.69
2014	Afghanistan	Ammonia, anhydrous	1.07
2017	Afghanistan	Ammonia, anhydrous	1.24
2002	Afghanistan	Ammonium nitrate (AN)	123.24
2007	Afghanistan	Ammonium nitrate (AN)	7.18

44 AsiaIV.describe()

	Year	Value
count	11480.000000	1.148000e+04
mean	2009.717073	3.102017e+04
std	4.570050	1.652045e+05
min	2002.000000	0.000000e+00
25%	2006.000000	5.258750e+01
50%	2010.000000	6.571850e+02
75%	2014.000000	6.525047e+03
max	2017.000000	6.294714e+06