گزارش crime in india

این دیتاست شامل ۷۶ تیبل مختلف از اطلاعات جرم و جنایت است.
تمامی جدول ها نشان دهنده اطلاعات مختلف در مناطق مختلف و سال های مختلف در هند است.

مناطق بر اساس ایالت ها و مناطق جزئی تر است و سال آن متغیر از ۲۰۰۱ تا ۲۰۱۳ است.
مثلا جدول زیر نشان دهنده اطلاعات اموال سرقت شده در هند در بازه ۲۰۰۱ تا ۲۰۱۰ بر اساس منطقه وقوع آن است.

	Area_Name	Year	Group_Name	Sub_Group_Name	Cases_Property_Recovered	Cases_Property_Stolen	Value_of_Property_Recovered	Value_of_Property_Stolen
0	Andaman & Nicobar Islands	2001	Burglary - Property	3. Burglary	27	64	755858	1321961
1	Andhra Pradesh	2001	Burglary - Property	3. Burglary	3321	7134	51483437	147019348
2	Arunachal Pradesh	2001	Burglary - Property	3. Burglary	66	248	825115	4931904
3	Assam	2001	Burglary - Property	3. Burglary	539	2423	3722850	21466955
4	Bihar	2001	Burglary - Property	3. Burglary	367	3231	2327135	17023937
...	...	...	...	...	...	...	...	...
2444	Tamil Nadu	2010	Total Property	7. Total Property Stolen & Recovered	16125	21509	660311804	1317919190
2445	Tripura	2010	Total Property	7. Total Property Stolen & Recovered	192	879	5666102	33032746
2446	Uttar Pradesh	2010	Total Property	7. Total Property Stolen & Recovered	9130	35068	577591772	1442670414
2447	Uttarakhand	2010	Total Property	7. Total Property Stolen & Recovered	964	2234	47135685	123398840
2448	West Bengal	2010	Total Property	7. Total Property Stolen & Recovered	4548	23759	1168242161	5015168687

2449 rows × 8 columns

برای به دست آوردن الگو در این دیتاها ما نیازمند این هستیم که دیتا ها را از این حالت پراکنده خارج کنیم.

مثلا ما اگر به جای ۷۶ جدول ۱ جدول داشتیم به سادگی می‌توانستیم کورلیشن ستون ها را با هم بررسی کنیم و میزان وابستگی دو فیچر را بفهمیم.
مثلا می‌خواهیم بفهمیم که آیا در مناطقی خشونت پلیس بیشتر باشد میزان قتل بیشتر است یا خیر.
برای اثبات این فرض می‌توانیم بررسی کنیم که مقدار کورلیشن خشونت پلیس و قتل در مناطق مختلف چقدر است و به این روش آماری آن را اثبات کنیم.
اما مشکل ما این است که اطلاعات مربوط به خشونت پلیس و قتل در جدول های جدایی هستند و برای محاسبه کرولیشن نیاز داریم که هر دو اطلاعات در یک ماتریس باشند.

پس هدف اصلی ما برای تمیز کردن دیتا ها این است که به نوعی بتوانیم همه جدول ها را با هم ادغام کنیم.
اما مشکل اصلی ای که با آن روبرو هستیم ثابت نبودن فرمت و ساختار جدول ها است. و اطلاعاتی مانند عنوان ستون مناطق و تعداد مناطقی که در آن اطلاعات موجود است و اطلاعاتی مانند این در جدول های مختلف به شکل متفاوتی آورده شده اند.

به طور مثال در جدول اموال سرقت شده عنوان ستون مناطق Area Name است و در جدول جرم های ثبت شده عنوان ستون STATE/UT است.

اولین کاری که سعی می‌کنیم انجام دهیم این است که جدول های مختلف را بر اساس مناطق اصلی (ایالت) ها گروه کنیم و مجموع ستون ها (که معمولا خودش یه دسته بندی جدا در ستون **subgroup** است) را نگه داریم.
و سپس بر اساس سال وقوع آن ها را فیلتر کنیم.
یعنی جدول های ما تبدیل می‌شوند به مجموع اطلاعات ثبت شده در ایالت ها در یک سال خاص
مثلا جدول زیر نشان دهنده مجموع اطلاعات همان جدول صفحه قبل در ایالت ها در سال ۲۰۰۱ است.

Index	Area_Name	Year	Group_Name	Sub_Group_Name	Cases_Property_Recovered	Cases_Property_Stolen	Value_of_Property_Recovered	Value_of_Property_Stolen
2100	Andaman & Nicobar Islands	2001	Total Property	7. Total Property Stolen & Recovered	54	143	1192179	3184477
2101	Andhra Pradesh	2001	Total Property	7. Total Property Stolen & Recovered	13418	25070	186103403	476038316
2102	Arunachal Pradesh	2001	Total Property	7. Total Property Stolen & Recovered	300	858	9652850	58483056
2103	Assam	2001	Total Property	7. Total Property Stolen & Recovered	2149	9778	24989343	121602215
2104	Bihar	2001	Total Property	7. Total Property Stolen & Recovered	3357	18503	47713186	422706220
2105	Chandigarh	2001	Total Property	7. Total Property Stolen & Recovered	714	1948	21114612	49527109
2106	Chhattisgarh	2001	Total Property	7. Total Property Stolen & Recovered	3298	9894	37331973	112242456
2107	Dadra & Nagar Haveli	2001	Total Property	7. Total Property Stolen & Recovered	43	106	5314436	11604547
2108	Daman & Diu	2001	Total Property	7. Total Property Stolen & Recovered	21	94	2323494	14151158
2109	Delhi	2001	Total Property	7. Total Property Stolen & Recovered	5893	25170	218254594	2127553393
2110	Goa	2001	Total Property	7. Total Property Stolen & Recovered	289	1004	9521526	239891783
2111	Gujarat	2001	Total Property	7. Total Property Stolen & Recovered	7385	24410	227693256	1149816876
2112	Haryana	2001	Total Property	7. Total Property Stolen & Recovered	5103	10536	257829983	404891438
2113	Himachal Pradesh	2001	Total Property	7. Total Property Stolen & Recovered	367	1573	24625652	59011695
2114	Jammu & Kashmir	2001	Total Property	7. Total Property Stolen & Recovered	833	3531	33575768	100502822
2115	Jharkhand	2001	Total Property	7. Total Property Stolen & Recovered	1411	5764	14052425	126668437
2116	Karnataka	2001	Total Property	7. Total Property Stolen & Recovered	6652	21039	218531303	583882867
2117	Kerala	2001	Total Property	7. Total Property Stolen & Recovered	3320	10442	99493533	433136856
2118	Lakshadweep	2001	Total Property	7. Total Property Stolen & Recovered	5	12	35400	109290
2119	Madhya Pradesh	2001	Total Property	7. Total Property Stolen & Recovered	11779	37554	203882247	501301044
2120	Maharashtra	2001	Total Property	7. Total Property Stolen & Recovered	20787	60173	856697065	4001169316
2121	Manipur	2001	Total Property	7. Total Property Stolen & Recovered	47	455	3658556	31017310
2122	Meghalaya	2001	Total Property	7. Total Property Stolen & Recovered	191	713	3662712	13868800
2123	Mizoram	2001	Total Property	7. Total Property Stolen & Recovered	866	1516	10038622	18857452
2124	Nagaland	2001	Total Property	7. Total Property Stolen & Recovered	120	463	6805838	81016077
2125	Odisha	2001	Total Property	7. Total Property Stolen & Recovered	5068	10167	52527851	220741872
2126	Puducherry	2001	Total Property	7. Total Property Stolen & Recovered	303	651	4421759	10439367
2127	Punjab	2001	Total Property	7. Total Property Stolen & Recovered	3045	5950	225006524	465830826
2128	Rajasthan	2001	Total Property	7. Total Property Stolen & Recovered	7854	25326	358218186	582726219
2129	Sikkim	2001	Total Property	7. Total Property Stolen & Recovered	40	129	1158845	3351263
2130	Tamil Nadu	2001	Total Property	7. Total Property Stolen & Recovered	17295	24606	333439558	735013778
2131	Tripura	2001	Total Property	7. Total Property Stolen & Recovered	130	579	724170	13759600
2132	Uttar Pradesh	2001	Total Property	7. Total Property Stolen & Recovered	9376	36711	846220887	1485857593
2133	Uttarakhand	2001	Total Property	7. Total Property Stolen & Recovered	549	2346	15855404	84480844
2134	West Bengal	2001	Total Property	7. Total Property Stolen & Recovered	4359	18204	111594745	427887883

حالا اگر بتوانیم تعدادی جدول را به این شکل در بیاوریم و ستون اطلاعات (در جدول بالا ستون های **case property recovered , case property stolen, value of property recovered, value of property recovered**) به یک جدول کلی که ایندکس های آن ایالت ها است اضافه کنیم می‌توانیم به نوعی اطلاعات جدول های مختلف را در کنار هم داشته باشیم.

برای این کار ابتدا اسم های ایالت ها را جدا می‌کنیم که در جدول کلی ایندکس قرار دهیم
در دیتا ست های مختلف از ۲۳ تا ۳۵ ایالت مختلف منحصر به فرد آمده است و ۳۵ ایالت را جدا میکنیم.
در جدول هایی که ایالت های کمتری داریم در ردیف ایالت هایی که موجود **null** قرار می‌دهیم.

هر جدول اطلاعات منحصر به فردی نسبت به جدول های دیگر دارد که با دادن آن اطلاعات به تابعی که نوشتیم می‌توانیم هر جدول را به شکل جدول بالا در بیاوریم و اطلاعات آن را به دیتاست اصلی و هدفمان اضافه کنیم.
این اطلاعات:

- عنوان ستون مناطق
- عنوان ستون زیر مجموعه مناطق
- اسمی که برای مجموع زیر مجموعه های هر منطقه در دیتاست آمده
- عنوان ستون سال
- سال مورد نظر
- لیست ستون هایی که لازم نیست به دیتاست اصلی اضافه شوند

به طور مثال در جدول مثال (اموال سرقت شده) اطلاعات به شکل زیر است

- Area Name
- Group Name
- Total Property
- Year
- 2001
- Area Name, Year, Group Name, Sub Group Name

اما در جدول جرم های ثبت شده اطلاعات به شکل زیر است

- STATE/UT
- DISTRICTS
- ZZ TOTAL
- YEAR
- 2001

نتایج

ما به این روش ترکیبات مختلفی از جدول های مختلف را با هم تست کردیم.
اما بهترین نتیجه نتیجه زیر بود

جدول هایی که با هم ادغام شدند:

10_Property_stolen_and_recovered.csv

35_Human_rights_violation_by_police.csv

01_District_wise_crimes_committed_IPC_2001_2012.csv

02_District_wise_crimes_committed_against_ST_2001_2012.csv

13_Police_killed_or_injured_on_duty.csv

جدول نهایی شامل ۶۱ ستون اطلاعات مختلف می‌باشد
و کورلیشن ستون ها را می توانیم در cell آخر نوت بوک مشاهده کنیم.

مثلا فیچر murder بعد از ستون attempt to murder که بدیهی است با فیچر policeman chargesheeted کورلیشن بیشتری دارد
می‌توانیم این اطلاعات را به دست بیاوریم در جاهایی که پلیس های بیشتری متهم شدند میزان قتل بیشتر است.