MACR Data Inventory

Abstract

The Monthly Arrest and Citations Record (MACR) database describes arrests in California from 1980 to the present. Each row includes the race, age, and gender of the individual arrested, the most serious offense for which he or she was arrested, the year of the arrest, and the county in which the arrest took place. This version of the data only contains arrests of adults (18+). The data have been de-identified using suppression and re-sampling, which means that any individual arrest may not appear in the data and counts may not match published statistics exactly.

Introduction

MACR Main Database

The Monthly Arrest and Citation Register compiles monthly arrest reports from all law enforcement agencies in California. The reports are supposed to contain every arrest (and citation until 2005) that an agency makes of juveniles or adults.

- Each row represents a single arrest, and includes the arrest date, the offense for which the individual was arrested, the individual's age, gender, race or ethnicity, and whether the case was referred to the local prosecutor. Each row also contains a full name and birthdate, which are excluded in these data.
- An arrest is defined as detaining an individual with the intention of seeking charges for a specific offense
- The MACR contains only the most serious offense based on the severity of possible punishment for retention.
- MACR data began to be stored in a digital format in 1980. The dataset spans 1980 to the present.

Law enforcement agencies are instructed to report all persons arrested within their jurisdiction. They are instructed to include arrests that result in a release without charges (including arrests of juveniles that only result in a warning). Departments that fail to send their data within thirty days of the due date are contacted, with increasing escalation when they reach 60 or 90 days past the due date.

Historical changes:

- The race_or_ethnicity codes for Asian/Pacific Islander expanded in 1991.
- The CA DOJ stopped collecting data about arrests or citations made for infractions in 2005.
- In 2011, the lower limit of felony theft was raised from $400 to $950, contributing to the decrease in felony theft arrests and increase in misdemeanor theft arrests.
- In 2011, some misdemeanor marijuana offenses were re-classified to infractions leading to a decrease in misdemeanor marijuana arrests.
- In 2014, California voters passed Proposition 47, which reduced numerous state statutes from felonies to misdemeanors - leading to a reduction in some types of felony arrests.

Department-specific changes:

- Bakersfield Police Department (PD) and Oakland PD did not report arrest data in 1995.
- San Francisco did not update its race_or_ethnicity codes until 2012, when it adoped the FBI's categories: white, black, American Indian, other Asian, and other. San Francisco data since 2012 does not distinguish between Hispanic and non-Hispanic whites.

Sample Rows

record_type_id	bcs_jurisdict...	ncic_jurisdic...	arrest_year	arrest_month	arrest_day	...
94	0	1900	1980	1	5	...
94	0	1900	1980	1	1	...
94	0	1900	1980	1	1	...
94	0	1900	1980	1	1	...
94	0	1900	1980	1	1	...
...	...	...	...	...	...	...

Variable Summary

name	type	value	description
age	integer	0-112	
arrest_day	integer	0-31	
arrest_month	integer	1-12	
arrest_year	integer	1980-2015	
bcs_jurisdiction	factor	0/1/5/7/12/28/36/40/45/...	deprecated
bcs_offense_code	factor	1/2/3/4/6/7/9/16/17/26/...	
bcs_summary_offense_code	factor	1/2/3/4/5/6/7/8/9/10/11...	
birth_day	integer	0-31	
birth_month	integer	0-12	
birth_year	integer	19-5010	
disposition	factor	released/turned ov.../m...	
fbi_offense_code	factor	01A/01B/02/03/04/05/06/...	
gender	factor	male/female	
name	character	pii	
id	integer	pii	
ncic_jurisdiction	factor	0100/0101/0102/0103/010...	
offense_level	factor	status of.../misdemean....	
race_or_ethnicity	factor	White/Hispanic/Black/Am...	
record_type_id	factor	14/21/24/32/91/94	
status_type	factor	cited/booked/other	
summary_offense_level	factor	felony/juvenile/misdeme...	

Not Shown

MACR, not yet cleaned

Sample Rows

record_type_id	bcs_jurisdict...	ncic_jurisdic...	arrest_year	arrest_month	arrest_day	...
94	0	1900	1980	1	5	...
94	0	1900	1980	1	1	...
94	0	1900	1980	1	1	...
94	0	1900	1980	1	1	...
94	0	1900	1980	1	1	...
...	...	...	...	...	...	...

BCS Code Table

BCS codes combine like statutes for statistical analysis. This table maps statutes to BCS codes to BCS summary codes (groups of BCS codes).

Sample Rows

offense_code	summary_offen...	summary_offen...	offense_categ...	new_2013
1	68	Truancy	Status	0
2	69	Runaway	Status	0
3	70	Curfew	Status	0
4	72	Other Stat Of...	Status	0
6	72	Other Stat Of...	Status	0
...	...	...	...	...

Variable Summary

name	type	value	description
offense_code	integer	1-998	
summary_offense_code	integer	1-76	groups offense codes
summary_offense_type	character	Truancy/Runaway/Curfew/...	
offense_category	character	Status/Misdemean.../Oth...	
new_2013	binary	0-1	law changed in 2013

NCIC Jurisdiction Table

The jurisdiction is the law enforcement agency that made the arrest. This table maps jurisdiction codes to their names and counties. It also describes when agencies began and stopped reporting, when agencies merged, and if agencies subcontracted to one another.

Sample Rows

CntyCode	County	Code	Agency	Start	End	...
1	Alameda County	0100	Alameda Co. S...			...
1	Alameda County	0101	Alameda			...
1	Alameda County	0102	Albany			...
1	Alameda County	0103	Berkeley			...
1	Alameda County	0104	Emeryville			...
...	...	...	...	...	...	...

Variable Summary

name	type	value	description
CntyCode	integer	1-58	
County	character	Alameda C.../Alpine Co....	
Code	character	0100/0101/0102/0103/010...	
Agency	character	Alameda C.../Alameda/Al...	

name	type	value	description
Start	character	/1/1/1997/1/1/2003/7/1/...	if absent, active throughout
End	character	/12/31/2003/6/30/2007/1...	
Contract	character	/C	
CJSC.Notes	character	/Name chan.../MACR only...	
Old.Juris.Code	character	20-000/20-002/20-004/20...	remove '-' to match MACR

Tables

Main Table Variables

age

Prompt

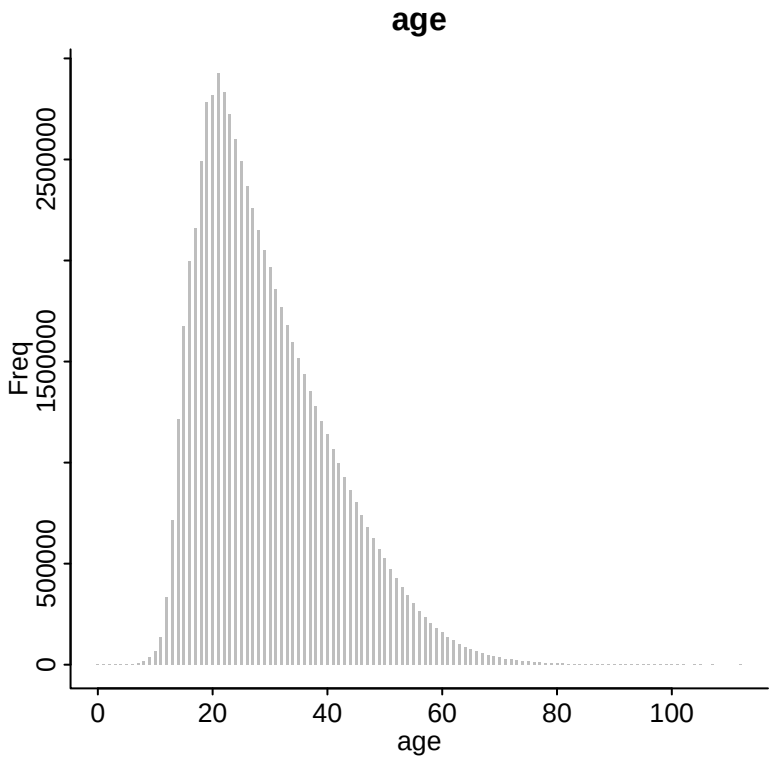
Arrest data for adults (age 18 years and older on the date of arrest) and juveniles (age 17 years or younger on the date of arrest) must be separated. Check the proper box to indicate if the data on the page submitted is adult or juvenile. If an agency has no adult or juvenile arrests for a month, "no adults to report" or "no juveniles to report" box must be checked.

Notes

if not already done so, users should consider dropping arrestees under age 5 and over age 89 as they may be data entry errors

Summary

Name	Value
Min.	0.00000
1st Qu.	21.00000
Median	27.00000
Mean	29.51242
3rd Qu.	36.00000
Max.	112.00000
NA's	76296.00000



arrest_day

Prompt

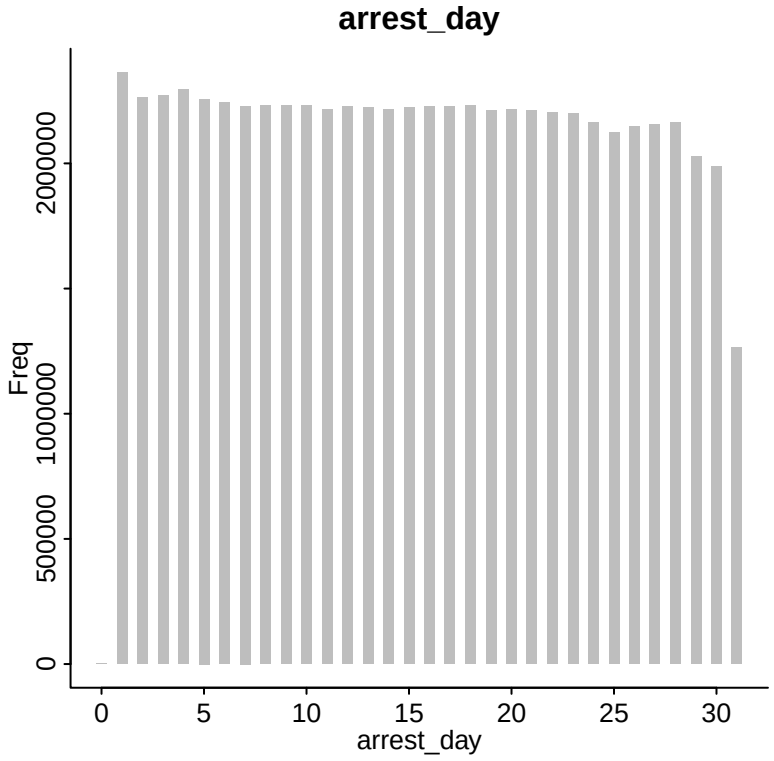
Use eight-digits: two each for the month, and day, and four for the year. For example, an arrest made on February 9, 2006 should be entered as: 02/09/2006.

Notes

The date February 30 ("02/30") was originally used to indicate a missing arrest date, these were recoded to NA

Summary

Name	Value
Min.	0.00000
1st Qu.	8.00000
Median	16.00000
Mean	15.56896
3rd Qu.	23.00000
Max.	31.00000



arrest_month

Prompt

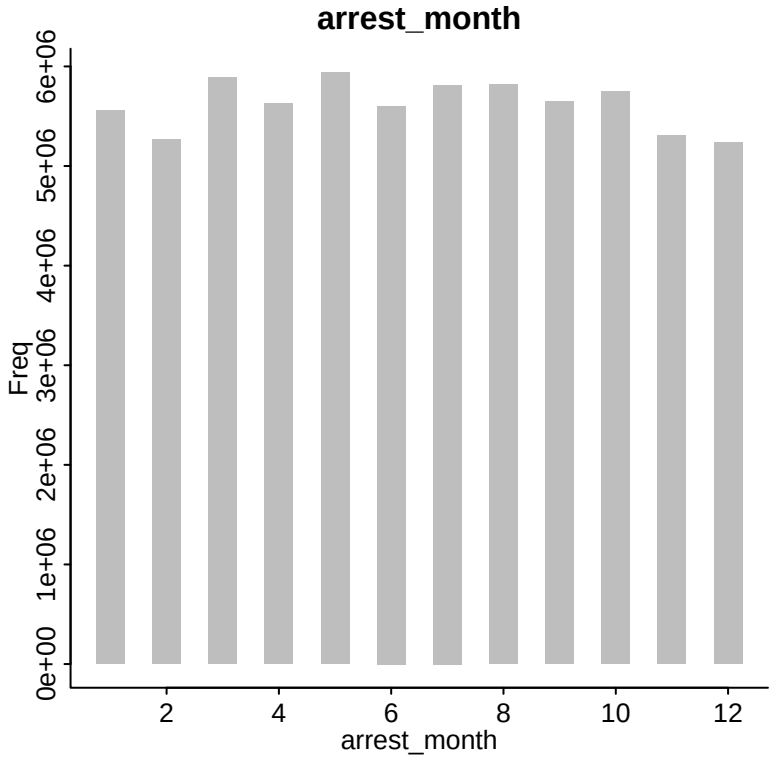
Use eight-digits: two each for the month, and day, and four for the year. For example, an arrest made on February 9, 2006 should be entered as: 02/09/2006.

Notes

The date February 30 ("02/30") was originally used to indicate a missing arrest date, these were recoded to NA

Summary

Name	Value
Min.	1.00000
1st Qu.	4.00000
Median	6.00000
Mean	6.46813
3rd Qu.	9.00000
Max.	12.00000



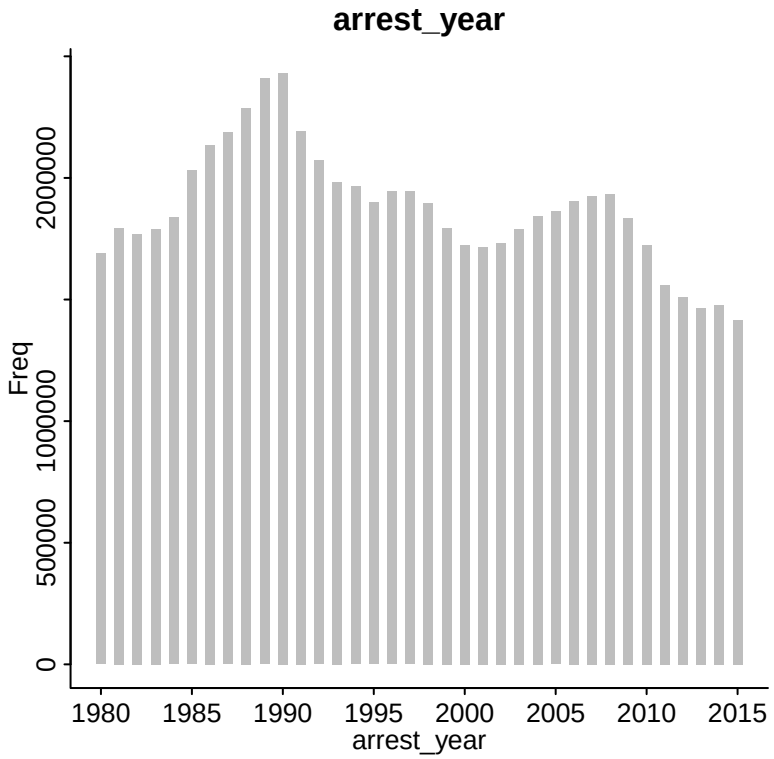
arrest_year

Prompt

Use eight-digits: two each for the month, and day, and four for the year. For example, an arrest made on February 9, 2006 should be entered as: 02/09/2006.

Summary

Name	Value
Min.	1980.000
1st Qu.	1988.000
Median	1996.000
Mean	1996.744
3rd Qu.	2005.000
Max.	2015.000



bcs_jurisdiction

Prompt

The code/ID number assigned to the reporting agency (old).

Notes

[should we remove this field?]

Summary

Name	Freq
NA	8181955
411	4370331
14664	2871932
21670	1754080
22682	1388334
40281	1374774
0	1238645
408	1108718
20512	960372
1	946241
14000	673520
54000	663168
54637	663102
45768	599052
11705	574494
12000	574174
46459	527164
13000	496708
41052	444986
12625	394901
41000	394344
13652	392131
16533	379916
11019	351185
...	...

bcs_offense_code

Labels

see the BCS offense codes table

Prompt

The code assigned to an offense. This code combines like statutes for statistical analysis.

Notes

This is the code for the most serious offense for which the individual was arrested. Officers and departments vary in their application of the penal code. By the time the code is entered onto the MACR, it might have been updated by detectives or by records clerks. [CJSC, is this true?: Depending on the knowledge and experience of the records clerks, coding consistency and accuracy might vary]. The BCS code system is maintanied by the Criminal Justice Statistics Center, with new penal codes being added as they gain usage. Misdemeanor traffic violations (BCS codes 086 and 087) are optional to report on the MACR

Summary

Name	Code	Freq
Drive Under the Influence	856	8449788
Misc Traffic	86	5814111
Drunk	46	5081866
Petty Theft	516	3432602
Failure to Appear/Non Traffic	98	2851950
Outside Warrant Misd	69	2780578
City/County Ordinance	97	2521050
Burglary	400	1995076
Other Drug Law Violations	836	1688291
Assault and Battery	397	1669267
Dangerous Drugs	825	1446113
Assault	372	1363754
Traffic	88	1359592
Narcotics	800	1240059
Marijuana	819	1161242
Other	96	1119441
Outside Warrant	65	903477
Assault	320	887030
Other Drug Law Violations	837	864658
Other Felony	993	851390
Liquor Laws	77	807106
Theft	530	797609
Motor Vehicle Theft	570	747147
Trespassing	68	714461
...	...	...

bcs_summary_offense_code

Labels

see the BCS offense codes table

Prompt

These codes are assigned to BCS codes. They combine like BCS codes for more general statistical analysis.

Notes

Because Summary Offense Codes combine BCS codes, the same caveats about potential coding errors at the individual or agency-level apply. Arrests for the following offenses are not included in publications from the California DOJ's Criminal Justice Statistics Center: * Summary code 26 = Felony Federal offense * Summary code 27 = Felony outside warrant * Summary code 28 = Felony probation/parole violation * Summary code 65 = Misdemeanor civil drunk * Summary code 66 = Misdemeanor outside warrant * Summary code 67 = Misdemeanor probation/parole violation * Summary code 74 = Misdemeanor miscellaneous traffic

Summary

Name	Code	Freq
Drive Under the Influence	51	8460894
Misc Traffic	74	5808841
Drunk	43	5081866
Petty Theft	31	3436020
Assault	6	3294714
Failure to Appear/Non Traffic	59	2851950
Assault and Battery	30	2815249
Outside Warrant Misd	66	2780578
Other Drug Law Violations	36	2676626
City/County Ordinance	58	2521050
Burglary	8	2263322
Narcotics	12	2027052
Dangerous Drugs	14	1895519
Theft	9	1883046
Other	60	1831728
Traffic	53	1727569
Other Felony	25	1459465
Marijuana	34	1315304
Liquor Laws	44	1022221
Motor Vehicle Theft	10	996304
Outside Warrant	27	903477
Robbery	5	812140
Trespassing	49	714476
Weapons	19	706417
...	...	...

birth_day

Prompt

Use eight-digits: two each for the month and day, and four for the year. For example, a birthdate of January 9, 1949 should be entered as: 01/09/1949. If the month and day are not known, use February 30 for the month and day and show the year of birth. For example, if the year of birth is 1945, enter the following: 02/30/1945. Do not write in the age. If the age is known, but not the date of birth, subtract the age from the present year and enter the resulting year of birth.

Notes

Because it was originally used to indicate missing values, “02/30” was recoded to NA

birth_month

Prompt

Use eight-digits: two each for the month and day, and four for the year. For example, a birthdate of January 9, 1949 should be entered as: 01/09/1949. If the month and day are not known, use February 30 for the month and day and show the year of birth. For example, if the year of birth is 1945, enter the following: 02/30/1945. Do not write in the age. If the age is known, but not the date of birth, subtract the age from the present year and enter the resulting year of birth.

Notes

Because it was originally used to indicate missing values, “02/30” was recoded to NA

birth_year

Prompt

Use eight-digits: two each for the month and day, and four for the year. For example, a birthdate of January 9, 1949 should be entered as: 01/09/1949. If the age is known, but not the date of birth, subtract the age from the present year and enter the resulting year of birth.

disposition

Labels

Misdemeanor (only for adults):

- Misdemeanor complaints that are sought by the arresting agency. (not used for juveniles)

Felony (only for adults):

- Felony complaints that are sought by the arresting agency. (not used for juveniles)

Released (only for adults):

- Each arrest released under 849(B) PC, or other sections, when no further action is planned by the arresting agency.
- Civil drunk arrest (647 (G) PC) or those individuals placed on other diversion programs by the local law enforcement agency, including those deemed not to be arrested.
- A new local offense in conjunction with an outside warrant. The level, status, charge, and disposition should be related to the local offense so that statistics on the local charges are captured. If the local offense is released so the out warrant may be acted upon, then the disposition is released.
- A new local offense in conjunction with a federal offense. The level, status, charge, and disposition should relate to the local offense so that statistics on the local charges are captured. If the local offense is released so the federal out warrant may be acted upon, then the disposition is released.
- not used for juveniles

Turned Over:

- Arrests made on another law enforcement agency's warrant (out warrant), with no local charges, and the subject is being held for the other agency
- Arrests made for a federal offense with no local charges.
- Fugitives from justice with no local charges.
- When a fine is paid to the local agency on a failure to appear traffic warrant issued by an outside jurisdiction and the money is forwarded to the issuing agency.

Juvenile Court:

- A juvenile that is referred to juvenile court or turned over to the probation department, welfare agency, other police agency, criminal or adult court or juvenile hall.

Department (only for juveniles):

- A situation that has been settled by the arresting agency, no action is to be taken by the juvenile probation department or the court, and the juvenile is released to his/her parents, guardian, or the street with a warning.
- A juvenile is placed on a local diversion program including, for statistical purposes, any juvenile deemed not arrested or cited.

Prompt

This column is intended for the disposition of the agency reporting the arrest or citation. DO NOT report the district attorney or court disposition in this column. It is intended to reflect the law enforcement agency disposition of the charge, not the person. ENTER ONE DISPOSITION PER LINE ITEM.

Summary

Name	Freq
misdemeanor complaint sought	37550031
felony complaint sought	12713879

Name	Freq
referred to juvenile probation department	6198209
turned over to other agency	4626341
released	4295910
handled within department	2009501
NA	76296

fbi_offense_code

Prompt

The FBI grouping of California offenses for national comparisons. These do not distinguish between felony and misdemeanor levels.

Notes

FBI codes only apply to the subset of offenses tracked in the Uniform Crime Reporting system

Summary

Name	Freq
NA	12049445
26	10672228
21	8683040
23	5085690
06	4347457
04	3271131
18E	3025619
08	2826141
18H	2543254
05	2297872
18F	1396677
15	1066642
07	1033748
22	1023003
14	955322
13	824370
03	812745
18A	794020
17	571353
16	565415
24	532292
18B	490897
18D	452739
28	438844
...	...

gender

Prompt

Enter either (1) Male or (2) Female.

Summary

Name	Freq
male	54724030
female	12669841
NA	76296

name

Prompt

Print legibly or type the last name, middle name or initial (if known), and first name of the arrestee. If name is unknown, use “John Doe” or “Jane Doe.”

id

Prompt

Enter the most reliable number for locating the arrested person in your agency's files in case questions arise. This can be the booking, arrest, or crime report number.

ncic_jurisdiction

Labels

see the NCIC jurisdiction table

Prompt

Enter your agency ORI/NCIC number. Agencies should abbreviate the nine-character NCIC code on the MACR report by using the fourth through seventh character of the NCIC code. For example, if your NCIC number is "CA0570100," report "5701" only.

Notes

Some agencies disappear and others are created over time. From 2005-2015, about 95% of arrests were made by about 250 of the 911 agencies in the dataset.

Summary

Name	Code	Freq
Los Angeles	1942	4370497
San Diego	3711	3019716
San Francisco	3801	1751020
Los Angeles Co. Sheriff's Department	1900	1696324
San Jose	4313	1424576
Fresno	1005	1374441
Long Beach	1941	1146733
Oakland	0109	1088790
CA Highway Patrol - Los Angeles	1999	946350
Sacramento	3404	880219
LAPD - Non-San Fernando Valley	193W	870135
Sacramento Co. Sheriff's Department	3400	816927
San Diego Co. Sheriff's Department	3700	715479
Bakersfield	1502	685932
San Bernardino Co. Sheriff's Department	3600	654412
Stockton	3905	610728
Santa Ana	3019	594182
Kern Co. Sheriff's Department	1500	589265
Riverside Co. Sheriff's Department	3300	574305
Modesto	5002	550950
Anaheim	3001	459238
Riverside	3313	414898
Oxnard	5604	410415
San Bernardino	3610	392134
...	...	...

offense_level

Labels

Status offense; Misdemeanor; Felony

Prompt

Select the level (delinquent, misdemeanor or felony) that best describes the most serious offense. Enter only one level per arrest or citation. 1) Delinquent (juvenile-only; also known as a status offense), 2) Misdemeanor, 3) Felony

Notes

Status offenses only apply to juveniles.

Summary

Name	Freq
misdemeanor	47383398
felony	18999927
status offense	1010546
NA	76296

race_or_ethnicity

Prompt

Record only one alpha designation that applies. Agencies submitting automated reports must verify that the appropriate codes are being entered. Do not report the race as "Unknown." Record the appropriate alpha code for race. Do not use "other" for unknown race.

Notes

The codes for Asian/Pacific Islander became more detailed in 1991. San Francisco did not change its reporting practices until 2012, when it adopted the FBI's categories for race: white, black, American Indian, other Asian, and other. Since 2012, San Francisco has not distinguished between non-Hispanic whites and Hispanic whites.

Summary

Name	Freq
White	27991111
Hispanic	24180079
Black	11763777
Other	1743674
Other Asian	391566
American Indian	376349
Filipino	293964
Vietnamese	149585
Chinese	122474
Pacific Islander	122024
NA	76296
Asian Indian	54884
Japanese	41810
Laotian	41506
Korean	35536
Hawaiian	29953
Samoan	29173
Cambodian	18532
Guamanian	7874

record_type_id

Labels

Arrest Codes:

- 14 - Add a record
- 24 - Replace a specific record
- 94 - Record sent to FBI

Records of No Arrest Codes:

- 21 - Report of no arrest
- 91 - Report of no arrest sent to FBI

Deleted Record Code:

- 32 - Specific delete action

Prompt

Flag that describes the action of the record. Codes 14, 24, and 94 represent arrest records. Codes 21, 32 and 91 represent deleted records or records of no arrest.

Summary

Name	Freq
94	65818499
14	1572148
32	76296
24	3224
21	0
91	0

status_type

Labels

Cited:

- Cited (or summoned) to appear in court as an alternative to being jailed or cited to court and later booked as directed by the court. A cite occurs in the field, when the suspect is not physically arrested by the officer.
- Informal booking -voluntarily go in and sign a notice to appear later in court.
- When a juvenile is cited in lieu of being delivered to juvenile authorities.

Booked:

- An adult is actually booked into jail for any period of time or booked into jail and later released on a citation.
- When a juvenile is booked into a juvenile holding facility of any type or any time an arrest report is filled out.

Other:

- An adult makes bail on a warrant and is neither cited nor booked.
- Detained for civil drunk occurrences per 647 (G) PC.
- When juveniles are neither cited nor booked (e.g., detained only, sent to a diversion program, referred to the probation department, etc.). Use "other" when there was no arrest report filled out.

Prompt

The status column describes the type of apprehension (at the time of initial contact with the arrestee). It determines how many individuals are cited versus those actually delivered to jail. The arresting agency is responsible for determining if it is a "cite," "book" or "other." The arresting agency should report "book" even when the suspect is sent to another law enforcement agency for processing. For example, many police departments send suspects that have been arrested to the county jail to be booked.

Notes

Booking rates vary to an implausible extent by agency and by year. Some agencies report 100% booking rates for every year. Other agencies report low booking rates for violent felonies. We recommend not using this variable unless you have reason to believe that particular agencies have reliable data.

Summary

Name	Freq
booked	47127834
cited	16548551
other	3717486
NA	76296

summary_offense_level

Labels

F - Felony (Adults) J - Juvenile M - Misdemeanor (Adults)

Prompt

The level distinguishs between juvenile and adult records.

Notes

"Juvenile" should match the count for those under 18.

Summary

Name	Freq
misdemeanor	42566580
felony	16458449
juvenile	8368842
NA	76296

Recommendations for Data Use

The MACR data are best used for analyses of general trends, they are less reliable for point estimates of numbers of arrests or numbers of people arrested. It is important to keep in mind that these data are heavily conditioned by individual, agency, and county variation in propensity to arrest, how offenses are categorized, and how well data are captured and reported to the CA DOJ. In using these data and preparing them for release, we have come across several anomalies and inconsistencies. They may produce results that are artefacts of data collection and reporting processes. To help researchers avoid potential pitfalls, we summarize our recommendations about data use below.

age

- Very young and very old ages are suspect. We suggest dropping those 5 or younger and 89 or older. VD NOTE: the problematic ages are as low as 80 in 1980

bcs_offense_code

- Arrest numbers for certain offenses may be more reliable than others. Some arrests, particularly for less serious offenses, may be missing. Different jurisdictions may report the same type of arrest using different codes.

bcs_summary_offense_code

- Arrest numbers for certain offenses may be more reliable than others. Some arrests, particularly for less serious offenses, may be missing. Different jurisdictions may report the same type of arrest using different codes.

county

- County totals may be affected by reporting irregularities, such as large drops in reported arrests in one jurisdiction. See the Variation in Number of Arrests section for an explanation and use the VarArrestsFlag indicator variable to keep track of jurisdictions or counties that may have been affected by reporting problems in a particular year.

disposition

- While these data appear to be overall reliable, we suggest analysts interested in particular counties examine disposition data by offense and year to ensure that trends appear reasonable.

ncic_jurisdiction

- Some jurisdictions have implausible data for certain years, such as a drop from a few hundred or a few thousand arrests to zero. See the Variation in Number of Arrests section for an explanation and use the VarArrestsFlag indicator variable to keep track of jurisdictions or counties that may have been affected by reporting problems in a particular year. Note that jurisdictions that report zero arrests in one year will not have any records in the data - they can be found by looking at trend data or at the List of Missing Jurisdiction-Years in the Variation in Number of Arrests section.

race_or_ethnicity

- More specific codes for Asian/Pacific Islander were added in 1991. Researchers may want to map these to a more general category.
- Post 2012, San Francisco does not count arrests of Hispanics separately. Researchers may want to treat San Francisco separately in addressing questions about race or ethnicity.

status_type

- Booking data appears to be unreliable overall. We recommend not using it.

Data Cleaning

Deleted Records

Records with a type id of 32 represent deleted rows, and as they contain no information about the kind of arrest made are deemed unusuable and omitted. A typical example would be:

reco...	bcs_...	ncic...	arre...	arre...	arre...	summ...	offe...	bcs_...	...
32	20000	0100	1996	6	14	NA	NA	NA	...
32	20000	0100	1996	6	29	NA	NA	NA	...
32	20000	0100	1996	6	30	NA	NA	NA	...
32	20000	0100	1996	6	1	NA	NA	NA	...

Arrest Date

A total of 880 records cannot be parsed into valid dates, i.e. the combination of `arrest_year`, `arrest_month`, and `arrest_day` results in a non-sensical date. Records with `arrest_day` of 0 are changed to `NA`, however in other cases it may the arrest day or month that causes the failure to parse. Some structure in the coding

may allow the correct date to be recovered. For example:

- Records at the end of a month with the `arrest_month` field prematurely incremented:

arrest_year	arrest_month	arrest_day	arrest_date	ncic_jurisdiction
1981	10	31	1981-10-31	4900
1981	10	31	1981-10-31	4900
1981	10	31	1981-10-31	4900
1981	11	31	NA	4900
1981	11	1	1981-11-01	4900
1981	11	1	1981-11-01	4900
1981	11	1	1981-11-01	4900

- Records where the digits in `arrest_day` appear to be transposed:

arrest_year	arrest_month	arrest_day	arrest_date	ncic_jurisdiction
1980	2	13	1980-02-13	1942
1980	2	13	1980-02-13	1942
1980	2	13	1980-02-13	1942
1980	2	31	NA	1942
1980	2	13	1980-02-13	1942
1980	2	13	1980-02-13	1942
1980	2	13	1980-02-13	1942

- Records where days were added at the end of a month:

arrest_year	arrest_month	arrest_day	arrest_date	ncic_jurisdiction
1980	2	29	1980-02-29	1942
1980	2	29	1980-02-29	1942
1980	2	29	1980-02-29	1942
1980	2	30	NA	1942
1980	2	30	NA	1942
1980	2	30	NA	1942
1980	2	31	NA	1942

As for now determining a 'correct arrest date' requires estimation, we leave those as is and augment the data with a field `arrest_date` that contains `NA` for all rows where a date cannot be parsed.

Birth Date

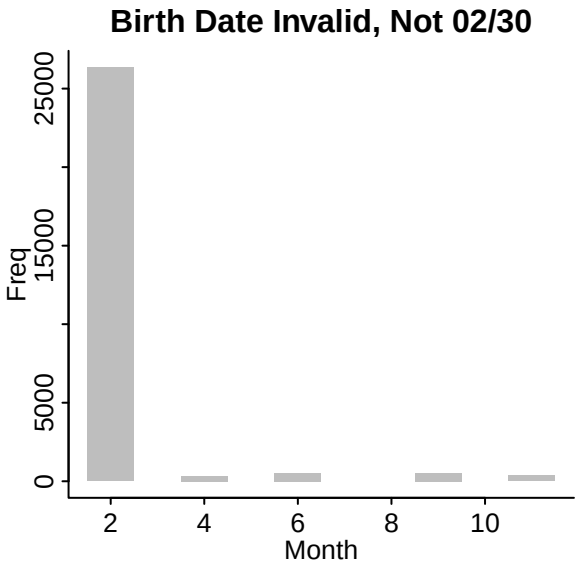
MACR includes fields for both the birth date and age, which do not always align. Errors in the birth date include:

- birth year recorded as `19xx` instead of `18xx`
- birth year recorded as `9xx` instead of `19xx`
- birth year in wrong millenia
- birth day 0
- birth month 0
- birth date invalid (e.g. `02/31/1991`)

Errors in the birth year are detected by looking for when the age and distance from arrest year to birth year exceed 1 in absolute value, and are 2372 in number. Cases where there difference is 100 or 1000 years can be resolved automatically, which leaves only 2 cases to be resolved by hand.

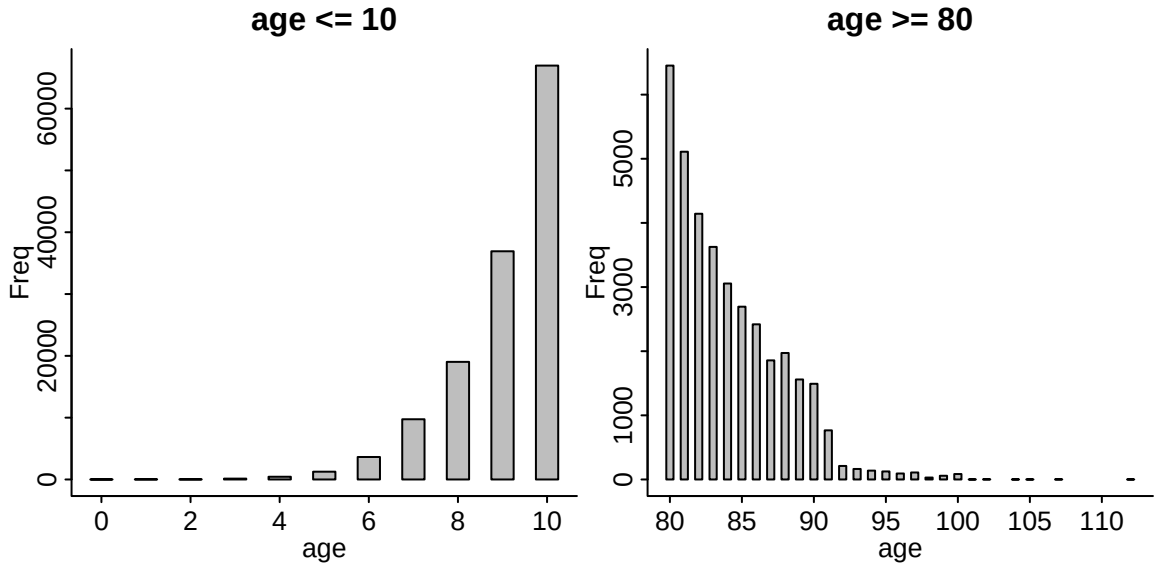
Of the remaining cases where a valid birth date can be derived from the birth year, month, and day, there are 74 cases where the distance from arrest date to birth date does not yield the recorded age. TODO: transition

The MACR manual states that if the specific birth date is unknown, the birth month and day should be recorded as February 30th. This accounts for 105452 of the remaining 133569 records with an invalid birth date. After excluding these records, there are still an anomalous amount of birth dates in February.



Variable Analysis

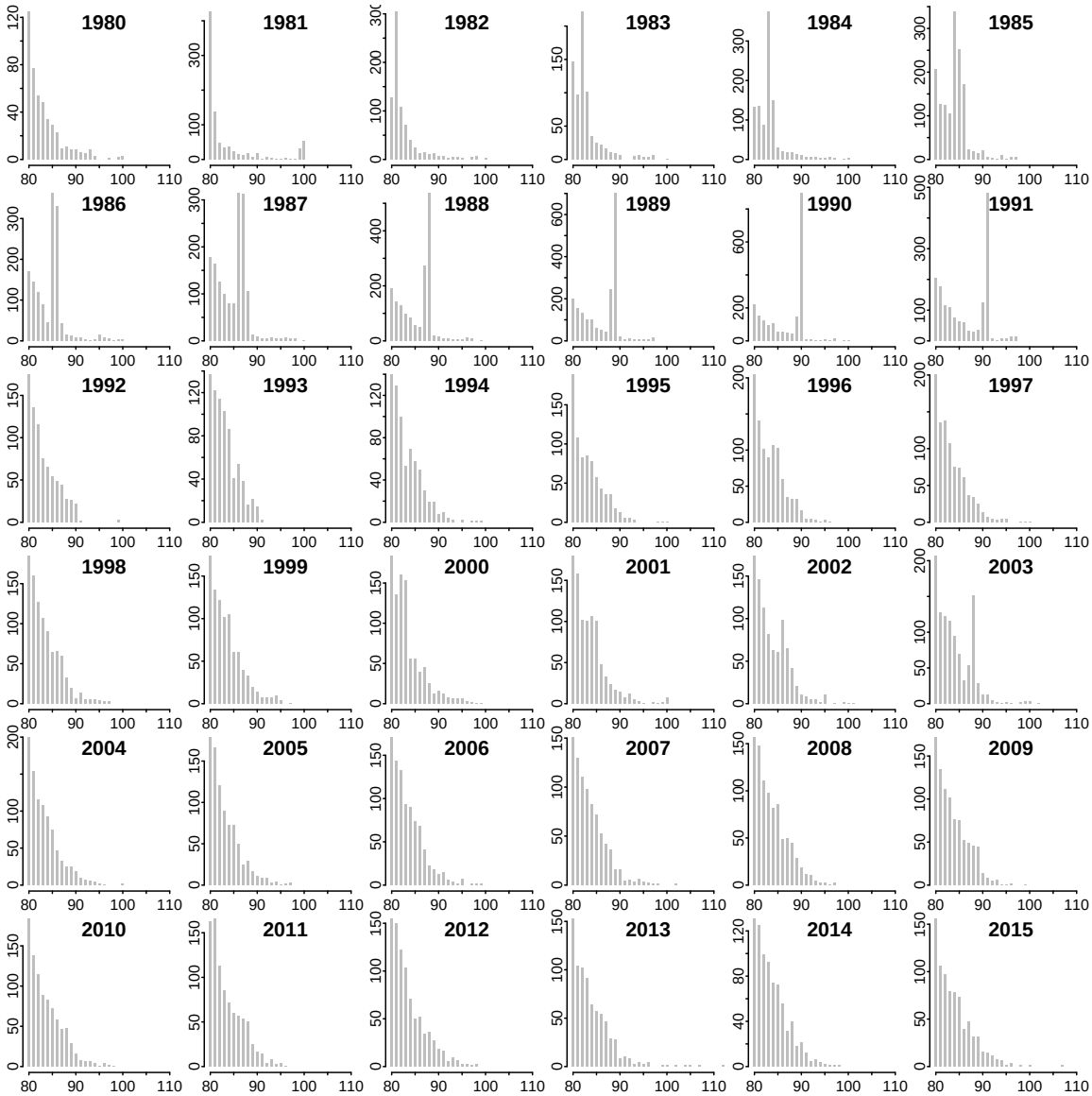
Age



As the above figure shows, there are a large number of arrests for exceptionally young children, and an odd decrease in arrests at age 90. For the young we find:

	age					
offense_level	0	1	2	3	4	5
status offense	1	11	10	36	68	139
misdemeanor	0	10	14	80	261	751
felony	3	8	3	24	113	369

Old age appears to be handled differently in different years. While the numbers are relatively small, it is difficult to believe that there were spikes in crime for, say, 91 year olds in 1991.



Directly examining these rows shows another form of missingness:

ncic_jurisdiction	arrest_date	age	birth_month	birth_day	birth_date
3710	1991-06-29	90	1	1	1901-01-01
3711	1991-11-07	91	1	1	1900-01-01
3310	1991-05-10	90	1	1	1901-01-01
3300	1991-06-25	90	1	1	1901-01-01
3300	1991-07-18	90	1	1	1901-01-01
3711	1991-11-01	91	1	1	1900-01-01
3700	1991-08-25	91	1	1	1900-01-01
3308	1991-05-03	89	10	2	1901-10-02
3300	1991-09-01	90	1	1	1901-01-01
3711	1991-11-06	91	1	1	1900-01-01
3300	1991-01-22	89	11	6	1901-11-06
5603	1991-09-06	90	1	1	1901-01-01
3709	1991-02-06	91	1	1	1900-01-01
3711	1991-12-04	91	1	1	1900-01-01
1947	1991-03-30	90	8	1	1900-08-01

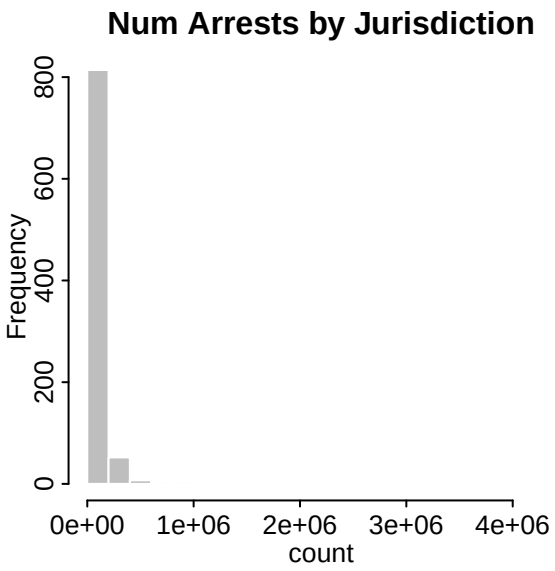
This spike apparently bubbles through the population until 1992, at which point the practice ended. In 2003, a handful of jurisidctions used an age of 88, together with a birthday of 02/30 to indicate missingness.

ncic_jurisdiction	arrest_date	age	birth_month	birth_day	birth_date
1005	2003-07-03	88	12	12	1914-12-12
1502	2003-08-30	88	1	20	1915-01-20
1900	2003-01-12	88	2	30	NA
1900	2003-01-26	88	2	30	NA
1900	2003-02-04	88	2	30	NA
1900	2003-02-11	88	2	30	NA
1900	2003-02-14	88	2	30	NA
1900	2003-02-14	88	2	30	NA
1900	2003-03-14	88	2	30	NA
1900	2003-03-16	88	2	30	NA

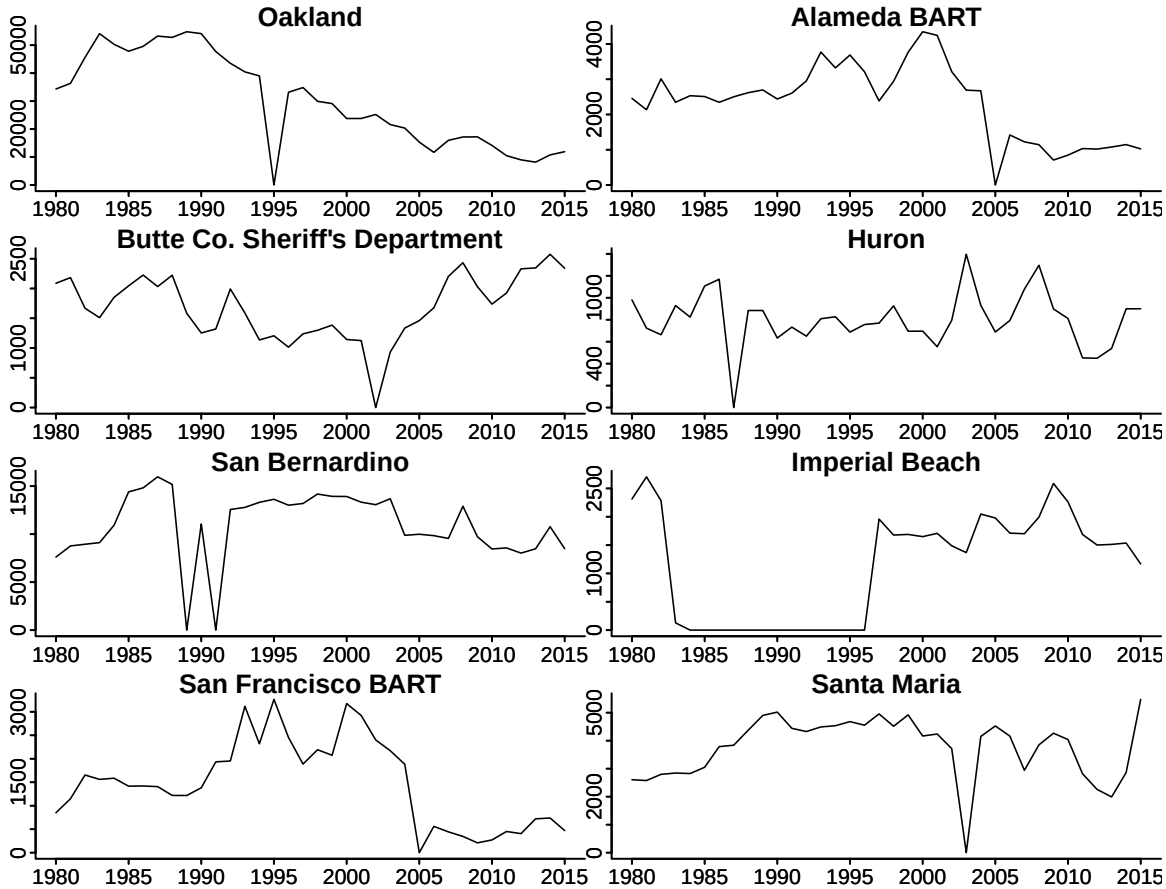
Recomendations

Jurisdictions

After removing deleted records, there are 889 different NCIC jurisdictions. The number of arrests in each varies wildly, from a minimum of 1 arrest across all 36 years to 4369571 arrests. The largest 396 account for 95% of the records, the smallest of which made 27464 arrests.



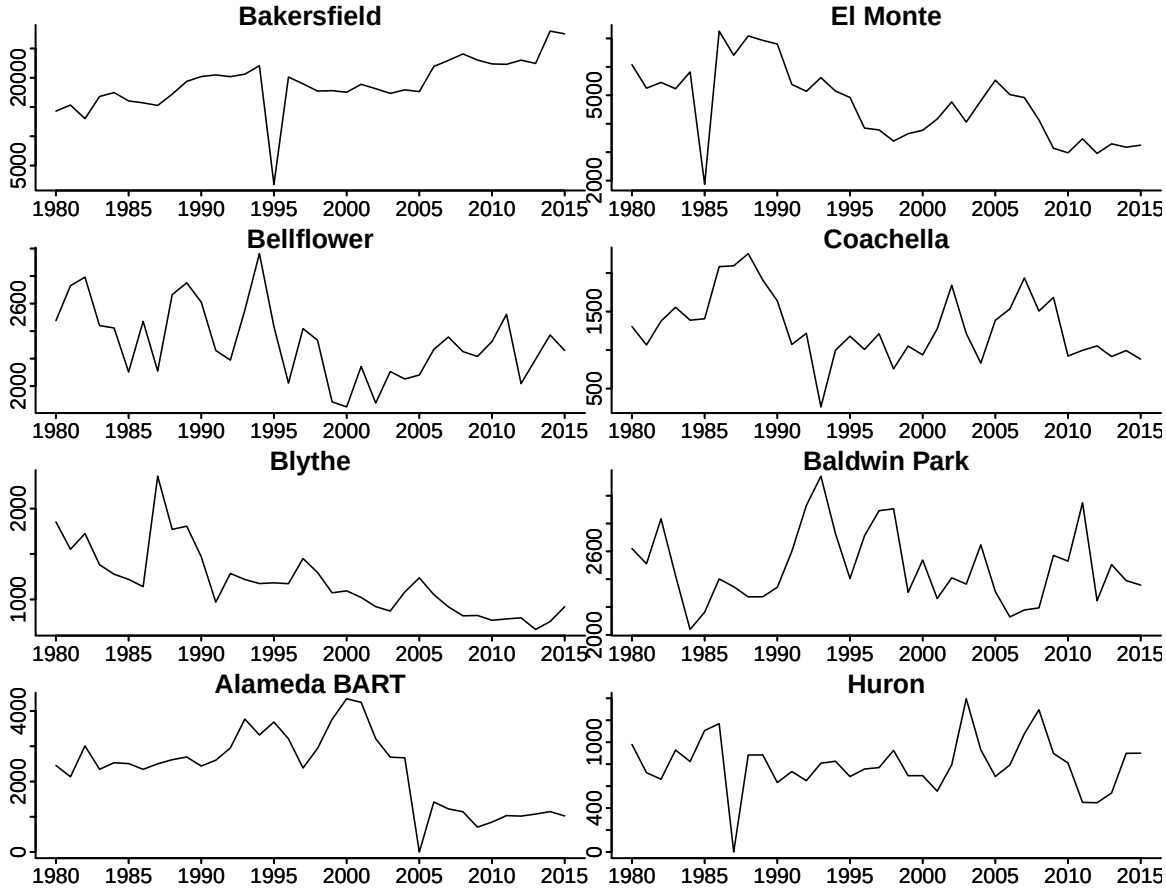
Furthermore, the number arrests within jurisdictions also appears to vary wildly over time. A total of 8 have unexpected years with 0 arrests, in some cases dropping from thousands of arrests to return to that rate immediately after.



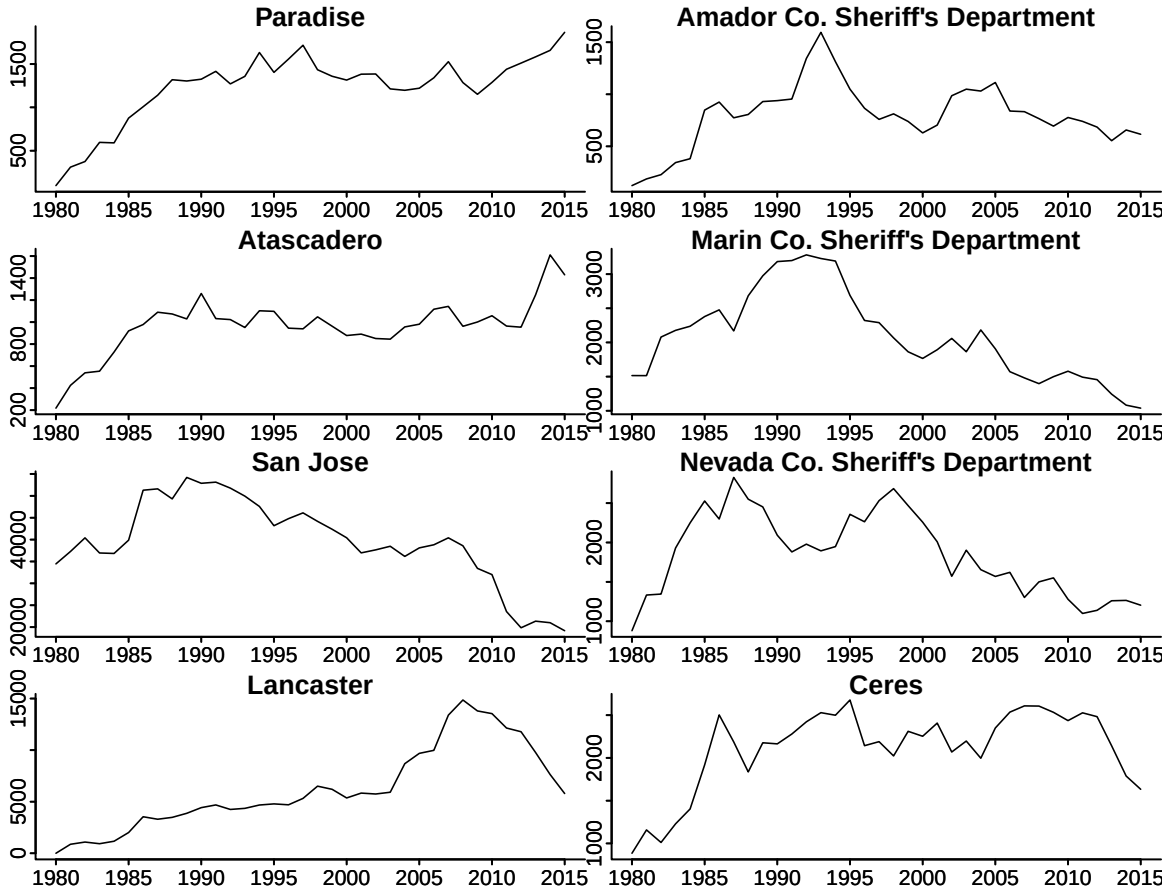
Variation in Number of Arrests

As a preliminary analysis, an hierarchical autoregressive model was fit to the number of arrests in each jurisdiction across time. The model included terms for how well the number of arrests in the previous year predicted the number of arrests in the next year (autoregressive coefficient), and terms for how much variability was shown by the jurisdictions after controlling for the average number of arrests. Also included, but not shown, were coefficients for the slopes which allowed jurisdictions to increase or decrease in their number of arrests over time. This is not an ideal model to fit to this kind of data, but should serve to capture the general trends.

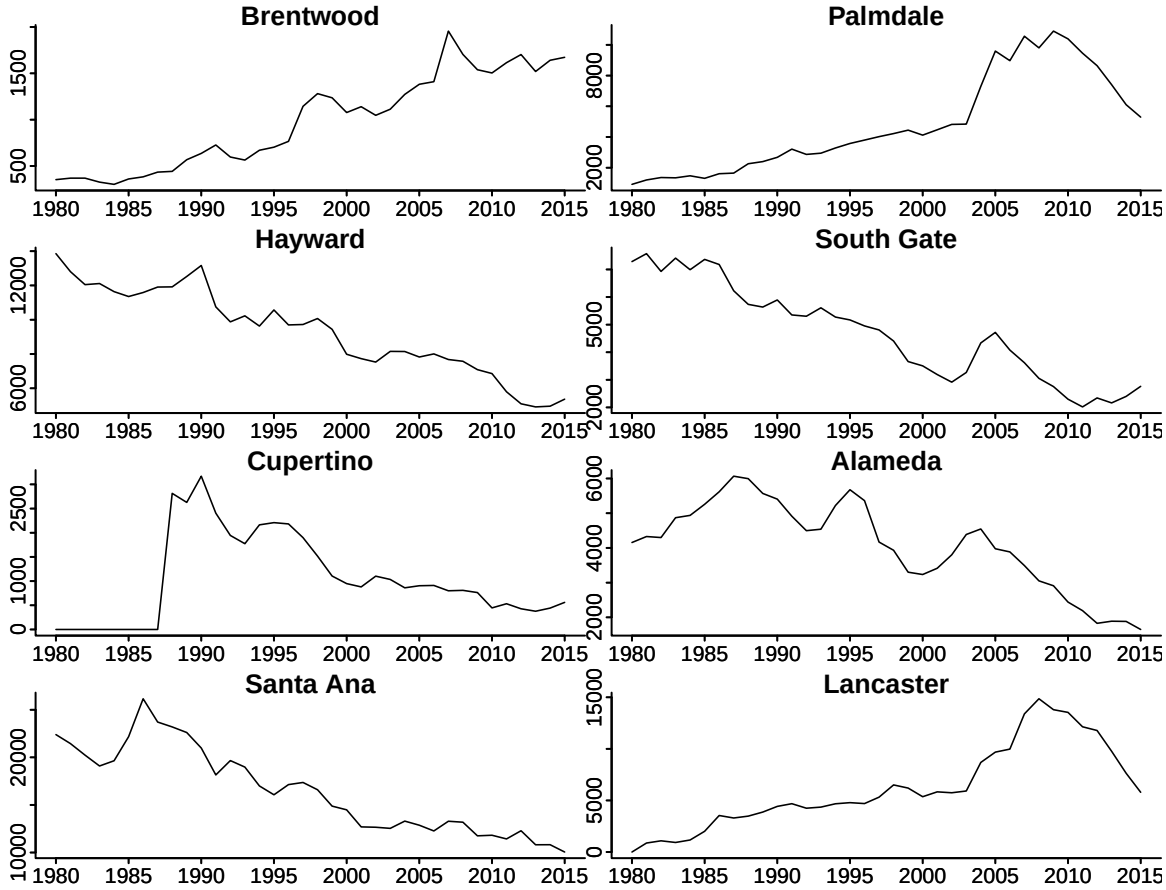
This graph shows those jursidictions with the lowest autoregression coefficients sorted from the lowest in the top left to the highest in the bottom right. A low autoregression essentially implies that knowing the previous year's value does not help in predicting the next, or that sometimes the next year's is much smaller or much larger and no pattern can be discerned. Because the model does not include the ability for the predicted line to bend (i.e. change points), large and sudden dips in the number of arrests cause it to fit poorly and are likely responsible for the worst cases presented.



Conversely, the next graph shows those jurisdictions with the highest autoregression coefficients, sorted with the highest in the top left and the lowest in the bottom right. Large values imply that the previous year's number of arrests was very often an excellent predictor of the current, and these jursidictions often have sequences of time where the number of arrests seems to move very little from an underlying predicted line.



Another form of variation in this context is how well the regression model predicts the observations, as reality should deviate just by chance. The following graph shows those jursidictions with extremely low variation with respect to what the model expects and with respect to the average number of arrests in that jursidiction. A jurisdiction that shows up here and not with a high autoregression coefficient would have the case that, from the previous year to the next, the number of arrests goes up or down randomly but overall, the number stays relatively close to a line.



Finally, we have those with extremely high levels of variation. Jurisdictions in this category either are simply noisy, or the straight-line plus autoregressive model is a poor fit.