

Data 100

Lecture 5: Data Cleaning & Exploratory Data Analysis

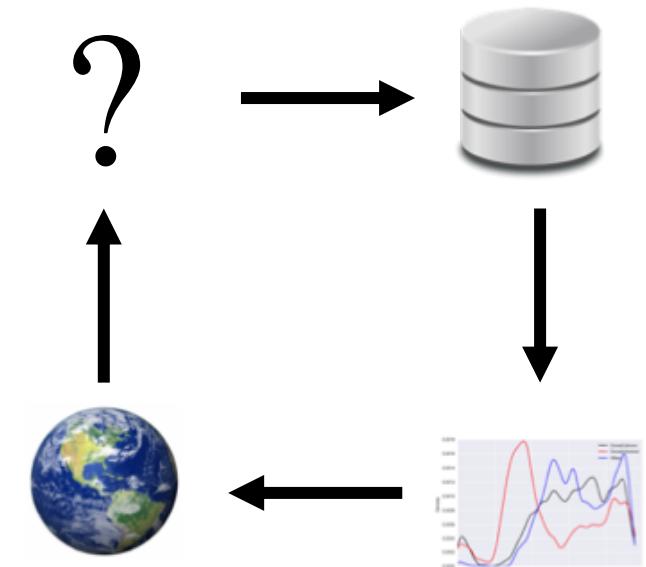
Slides by:

Joseph E. Gonzalez, Deb Nolan, & Joe Hellerstein

jegonzal@berkeley.edu

deborah_nolan@berkeley.edu

hellerstein@berkeley.edu



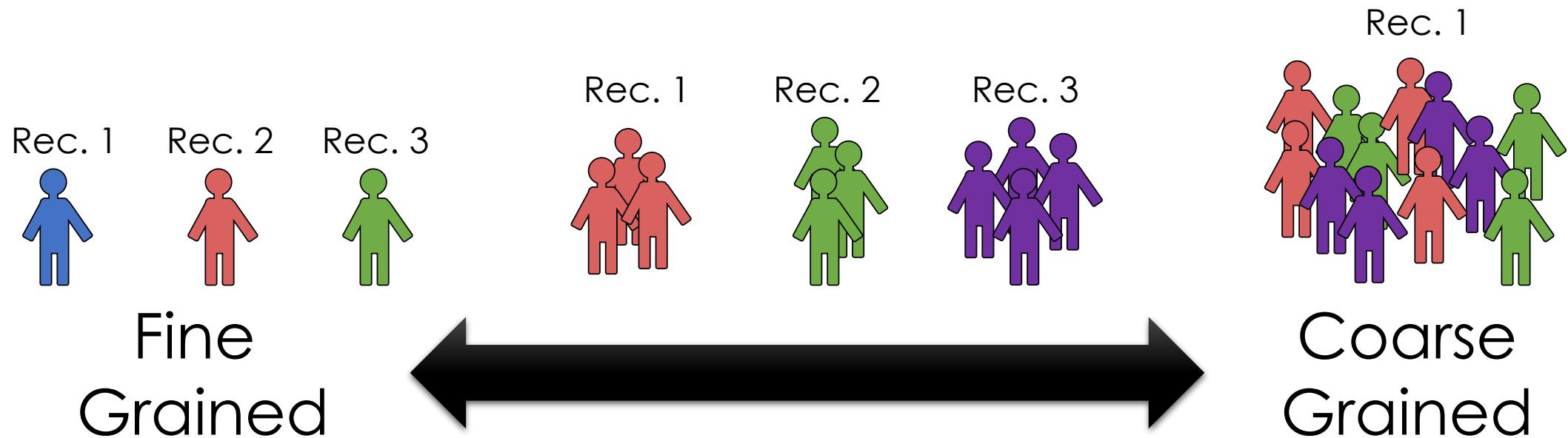
Last Lecture

- Started discussing exploratory data analysis
- **Structure** -- the “shape” of a data file (how is it organized)

```
1 CASENO OFFENSE EVENTDT EVENTTM CVLEGEND CVDOW IndoDate Block_Location
2 BLKADDR City State
2 18000273 VEHICLE STOLEN 01/01/2018 12:00:00 AM 20:30 MOTOR VEHICLE THEFT
1 01/24/2018 03:30:18 AM "1100 PARKER ST
3 Berkeley CA
4 (37.85,-122.288914)
5 178915: 1 CASENO,OFFENSE,EVENTDT,EVENTTM,CVLEGEND,CVDOW,IndoDate,Block_Location,BLKADDR,City,Stat
2 01
6 Berkeley CA
7 (37.85,-122.288914)
8 178925: 3 (37.859364, -122.288914)",1100 PARKER ST,Berkeley,CA
9 01
10 5 17892476, BURGLARY AUTO,12/12/2017 12:00:00 AM,13:30,BURGLARY - VEHICLE,2,01/24/2018
11 03:30:17 AM,"2300 LE CONTE AVE
12 (37.85,-122.288914)
13 178915: 7 (3
14 01
15 2
16 03
17 2
18 03
19 3
20 4
21 5
22 6
23 7
24 8
25 9
26 10
27 11
28 12
29 13
30 14
31 15
32 16
33 17
34 18
35 19
36 20
37 21
38 22
39 23
40 24
41 25
42 26
43 27
44 28
45 29
46 30
47 31
48 32
49 33
50 34
51 35
52 36
53 37
54 38
55 39
56 40
57 41
58 42
59 43
60 44
61 45
62 46
63 47
64 48
65 49
66 50
67 51
68 52
69 53
70 54
71 55
72 56
73 57
74 58
75 59
76 60
77 61
78 62
79 63
80 64
81 65
82 66
83 67
84 68
85 69
86 70
87 71
88 72
89 73
90 74
91 75
92 76
93 77
94 78
95 79
96 80
97 81
98 82
99 83
100 84
101 85
102 86
103 87
104 88
105 89
106 90
107 91
108 92
109 93
110 94
111 95
112 96
113 97
114 98
115 99
116 100
117 101
118 102
119 103
120 104
121 105
122 106
123 107
124 108
125 109
126 110
127 111
128 112
129 113
130 114
131 115
132 116
133 117
134 118
135 119
136 120
137 121
138 122
139 123
140 124
141 125
142 126
143 127
144 128
145 129
146 130
147 131
148 132
149 133
150 134
151 135
152 136
153 137
154 138
155 139
156 140
157 141
158 142
159 143
160 144
161 145
162 146
163 147
164 148
165 149
166 150
167 151
168 152
169 153
170 154
171 155
172 156
173 157
174 158
175 159
176 160
177 161
178 162
179 163
180 164
181 165
182 166
183 167
184 168
185 169
186 170
187 171
188 172
189 173
190 174
191 175
192 176
193 177
194 178
195 179
196 180
197 181
198 182
199 183
200 184
201 185
202 186
203 187
204 188
205 189
206 190
207 191
208 192
209 193
210 194
211 195
212 196
213 197
214 198
215 199
216 200
217 201
218 202
219 203
220 204
221 205
222 206
223 207
224 208
225 209
226 210
227 211
228 212
229 213
230 214
231 215
232 216
233 217
234 218
235 219
236 220
237 221
238 222
239 223
240 224
241 225
242 226
243 227
244 228
245 229
246 230
247 231
248 232
249 233
250 234
251 235
252 236
253 237
254 238
255 239
256 240
257 241
258 242
259 243
260 244
261 245
262 246
263 247
264 248
265 249
266 250
267 251
268 252
269 253
270 254
271 255
272 256
273 257
274 258
275 259
276 260
277 261
278 262
279 263
280 264
281 265
282 266
283 267
284 268
285 269
286 270
287 271
288 272
289 273
290 274
291 275
292 276
293 277
294 278
295 279
296 280
297 281
298 282
299 283
300 284
301 285
302 286
303 287
304 288
305 289
306 290
307 291
308 292
309 293
310 294
311 295
312 296
313 297
314 298
315 299
316 300
317 301
318 302
319 303
320 304
321 305
322 306
323 307
324 308
325 309
326 310
327 311
328 312
329 313
330 314
331 315
332 316
333 317
334 318
335 319
336 320
337 321
338 322
339 323
340 324
341 325
342 326
343 327
344 328
345 329
346 330
347 331
348 332
349 333
350 334
351 335
352 336
353 337
354 338
355 339
356 340
357 341
358 342
359 343
360 344
361 345
362 346
363 347
364 348
365 349
366 350
367 351
368 352
369 353
370 354
371 355
372 356
373 357
374 358
375 359
376 360
377 361
378 362
379 363
380 364
381 365
382 366
383 367
384 368
385 369
386 370
387 371
388 372
389 373
390 374
391 375
392 376
393 377
394 378
395 379
396 380
397 381
398 382
399 383
400 384
401 385
402 386
403 387
404 388
405 389
406 390
407 391
408 392
409 393
410 394
411 395
412 396
413 397
414 398
415 399
416 400
417 401
418 402
419 403
420 404
421 405
422 406
423 407
424 408
425 409
426 410
427 411
428 412
429 413
430 414
431 415
432 416
433 417
434 418
435 419
436 420
437 421
438 422
439 423
440 424
441 425
442 426
443 427
444 428
445 429
446 430
447 431
448 432
449 433
450 434
451 435
452 436
453 437
454 438
455 439
456 440
457 441
458 442
459 443
460 444
461 445
462 446
463 447
464 448
465 449
466 450
467 451
468 452
469 453
470 454
471 455
472 456
473 457
474 458
475 459
476 460
477 461
478 462
479 463
480 464
481 465
482 466
483 467
484 468
485 469
486 470
487 471
488 472
489 473
490 474
491 475
492 476
493 477
494 478
495 479
496 480
497 481
498 482
499 483
500 484
501 485
502 486
503 487
504 488
505 489
506 490
507 491
508 492
509 493
510 494
511 495
512 496
513 497
514 498
515 499
516 500
517 501
518 502
519 503
520 504
521 505
522 506
523 507
524 508
525 509
526 510
527 511
528 512
529 513
530 514
531 515
532 516
533 517
534 518
535 519
536 520
537 521
538 522
539 523
540 524
541 525
542 526
543 527
544 528
545 529
546 530
547 531
548 532
549 533
550 534
551 535
552 536
553 537
554 538
555 539
556 540
557 541
558 542
559 543
560 544
561 545
562 546
563 547
564 548
565 549
566 550
567 551
568 552
569 553
570 554
571 555
572 556
573 557
574 558
575 559
576 560
577 561
578 562
579 563
580 564
581 565
582 566
583 567
584 568
585 569
586 570
587 571
588 572
589 573
590 574
591 575
592 576
593 577
594 578
595 579
596 580
597 581
598 582
599 583
600 584
601 585
602 586
603 587
604 588
605 589
606 590
607 591
608 592
609 593
610 594
611 595
612 596
613 597
614 598
615 599
616 600
617 601
618 602
619 603
620 604
621 605
622 606
623 607
624 608
625 609
626 610
627 611
628 612
629 613
630 614
631 615
632 616
633 617
634 618
635 619
636 620
637 621
638 622
639 623
640 624
641 625
642 626
643 627
644 628
645 629
646 630
647 631
648 632
649 633
650 634
651 635
652 636
653 637
654 638
655 639
656 640
657 641
658 642
659 643
660 644
661 645
662 646
663 647
664 648
665 649
666 650
667 651
668 652
669 653
670 654
671 655
672 656
673 657
674 658
675 659
676 660
677 661
678 662
679 663
680 664
681 665
682 666
683 667
684 668
685 669
686 670
687 671
688 672
689 673
690 674
691 675
692 676
693 677
694 678
695 679
696 680
697 681
698 682
699 683
700 684
701 685
702 686
703 687
704 688
705 689
706 690
707 691
708 692
709 693
710 694
711 695
712 696
713 697
714 698
715 699
716 700
717 701
718 702
719 703
720 704
721 705
722 706
723 707
724 708
725 709
726 710
727 711
728 712
729 713
730 714
731 715
732 716
733 717
734 718
735 719
736 720
737 721
738 722
739 723
740 724
741 725
742 726
743 727
744 728
745 729
746 730
747 731
748 732
749 733
750 734
751 735
752 736
753 737
754 738
755 739
756 740
757 741
758 742
759 743
760 744
761 745
762 746
763 747
764 748
765 749
766 750
767 751
768 752
769 753
770 754
771 755
772 756
773 757
774 758
775 759
776 760
777 761
778 762
779 763
780 764
781 765
782 766
783 767
784 768
785 769
786 770
787 771
788 772
789 773
790 774
791 775
792 776
793 777
794 778
795 779
796 780
797 781
798 782
799 783
800 784
801 785
802 786
803 787
804 788
805 789
806 790
807 791
808 792
809 793
810 794
811 795
812 796
813 797
814 798
815 799
816 800
817 801
818 802
819 803
820 804
821 805
822 806
823 807
824 808
825 809
826 810
827 811
828 812
829 813
830 814
831 815
832 816
833 817
834 818
835 819
836 820
837 821
838 822
839 823
840 824
841 825
842 826
843 827
844 828
845 829
846 830
847 831
848 832
849 833
850 834
851 835
852 836
853 837
854 838
855 839
856 840
857 841
858 842
859 843
860 844
861 845
862 846
863 847
864 848
865 849
866 850
867 851
868 852
869 853
870 854
871 855
872 856
873 857
874 858
875 859
876 860
877 861
878 862
879 863
880 864
881 865
882 866
883 867
884 868
885 869
886 870
887 871
888 872
889 873
890 874
891 875
892 876
893 877
894 878
895 879
896 880
897 881
898 882
899 883
900 884
901 885
902 886
903 887
904 888
905 889
906 890
907 891
908 892
909 893
910 894
911 895
912 896
913 897
914 898
915 899
916 900
917 901
918 902
919 903
920 904
921 905
922 906
923 907
924 908
925 909
926 910
927 911
928 912
929 913
930 914
931 915
932 916
933 917
934 918
935 919
936 920
937 921
938 922
939 923
940 924
941 925
942 926
943 927
944 928
945 929
946 930
947 931
948 932
949 933
950 934
951 935
952 936
953 937
954 938
955 939
956 940
957 941
958 942
959 943
960 944
961 945
962 946
963 947
964 948
965 949
966 950
967 951
968 952
969 953
970 954
971 955
972 956
973 957
974 958
975 959
976 960
977 961
978 962
979 963
980 964
981 965
982 966
983 967
984 968
985 969
986 970
987 971
988 972
989 973
990 974
991 975
992 976
993 977
994 978
995 979
996 980
997 981
998 982
999 983
1000 984
1001 985
1002 986
1003 987
1004 988
1005 989
1006 990
1007 991
1008 992
1009 993
1010 994
1011 995
1012 996
1013 997
1014 998
1015 999
1016 1000
1017 1001
1018 1002
1019 1003
1020 1004
1021 1005
1022 1006
1023 1007
1024 1008
1025 1009
1026 1010
1027 1011
1028 1012
1029 1013
1030 1014
1031 1015
1032 1016
1033 1017
1034 1018
1035 1019
1036 1020
1037 1021
1038 1022
1039 1023
1040 1024
1041 1025
1042 1026
1043 1027
1044 1028
1045 1029
1046 1030
1047 1031
1048 1032
1049 1033
1050 1034
1051 1035
1052 1036
1053 1037
1054 1038
1055 1039
1056 1040
1057 1041
1058 1042
1059 1043
1060 1044
1061 1045
1062 1046
1063 1047
1064 1048
1065 1049
1066 1050
1067 1051
1068 1052
1069 1053
1070 1054
1071 1055
1072 1056
1073 1057
1074 1058
1075 1059
1076 1060
1077 1061
1078 1062
1079 1063
1080 1064
1081 1065
1082 1066
1083 1067
1084 1068
1085 1069
1086 1070
1087 1071
1088 1072
1089 1073
1090 1074
1091 1075
1092 1076
1093 1077
1094 1078
1095 1079
1096 1080
1097 1081
1098 1082
1099 1083
1100 1084
1101 1085
1102 1086
1103 1087
1104 1088
1105 1089
1106 1090
1107 1091
1108 1092
1109 1093
1110 1094
1111 1095
1112 1096
1113 1097
1114 1098
1115 1099
1116 1100
1117 1101
1118 1102
1119 1103
1120 1104
1121 1105
1122 1106
1123 1107
1124 1108
1125 1109
1126 1110
1127 1111
1128 1112
1129 1113
1130 1114
1131 1115
1132 1116
1133 1117
1134 1118
1135 1119
1136 1120
1137 1121
1138 1122
1139 1123
1140 1124
1141 1125
1142 1126
1143 1127
1144 1128
1145 1129
1146 1130
1147 1131
1148 1132
1149 1133
1150 1134
1151 1135
1152 1136
1153 1137
1154 1138
1155 1139
1156 1140
1157 1141
1158 1142
1159 1143
1160 1144
1161 1145
1162 1146
1163 1147
1164 1148
1165 1149
1166 1150
1167 1151
1168 1152
1169 1153
1170 1154
1171 1155
1172 1156
1173 1157
1174 1158
1175 1159
1176 1160
1177 1161
1178 1162
1179 1163
1180 1164
1181 1165
1182 1166
1183 1167
1184 1168
1185 1169
1186 1170
1187 1171
1188 1172
1189 1173
1190 1174
1191 1175
1192 1176
1193 1177
1194 1178
1195 1179
1196 1180
1197 1181
1198 1182
1199 1183
1200 1184
1201 1185
1202 1186
1203 1187
1204 1188
1205 1189
1206 1190
1207 1191
1208 1192
1209 1193
1210 1194
1211 1195
1212 1196
1213 1197
1214 1198
1215 1199
1216 1200
1217 1201
1218 1202
1219 1203
1220 1204
1221 1205
1222 1206
1223 1207
1224 1208
1225 1209
1226 1210
1227 1211
1228 1212
1229 1213
1230 1214
1231 1215
1232 1216
1233 1217
1234 1218
1235 1219
1236 1220
1237 1221
1238 1222
1239 1223
1240 1224
1241 1225
1242 1226
1243 1227
1244 1228
1245 1229
1246 1230
1247 1231
1248 1232
1249 1233
1250 1234
1251 1235
1252 1236
1253 1237
1254 1238
1255 1239
1256 1240
1257 1241
1258 1242
1259 1243
1260 1244
1261 1245
1262 1246
1263 1247
1264 1248
1265 1249
1266 1250
1267 1251
1268 1252
1269 1253
1270 1254
1271 1255
1272 1256
1273 1257
1274 1258
1275 1259
1276 1260
1277 1261
1278 1262
1279 1263
1280 1264
1281 1265
1282 1266
1283 1
```

Last Lecture

- Started discussing exploratory data analysis
- **Structure** -- *the “shape” of a data file (how is it organized)*
- **Granularity** -- *how fine/coarse is each datum*



Group By – manipulating granularity

Key Data

A	3
B	1
C	4
A	1
B	5
C	9
A	2
B	6
C	5

Split into Groups

A	3
B	1
C	9
A	2
B	6
C	5

Aggregate Function

A	6
---	---

Aggregate Function

B	12
---	----

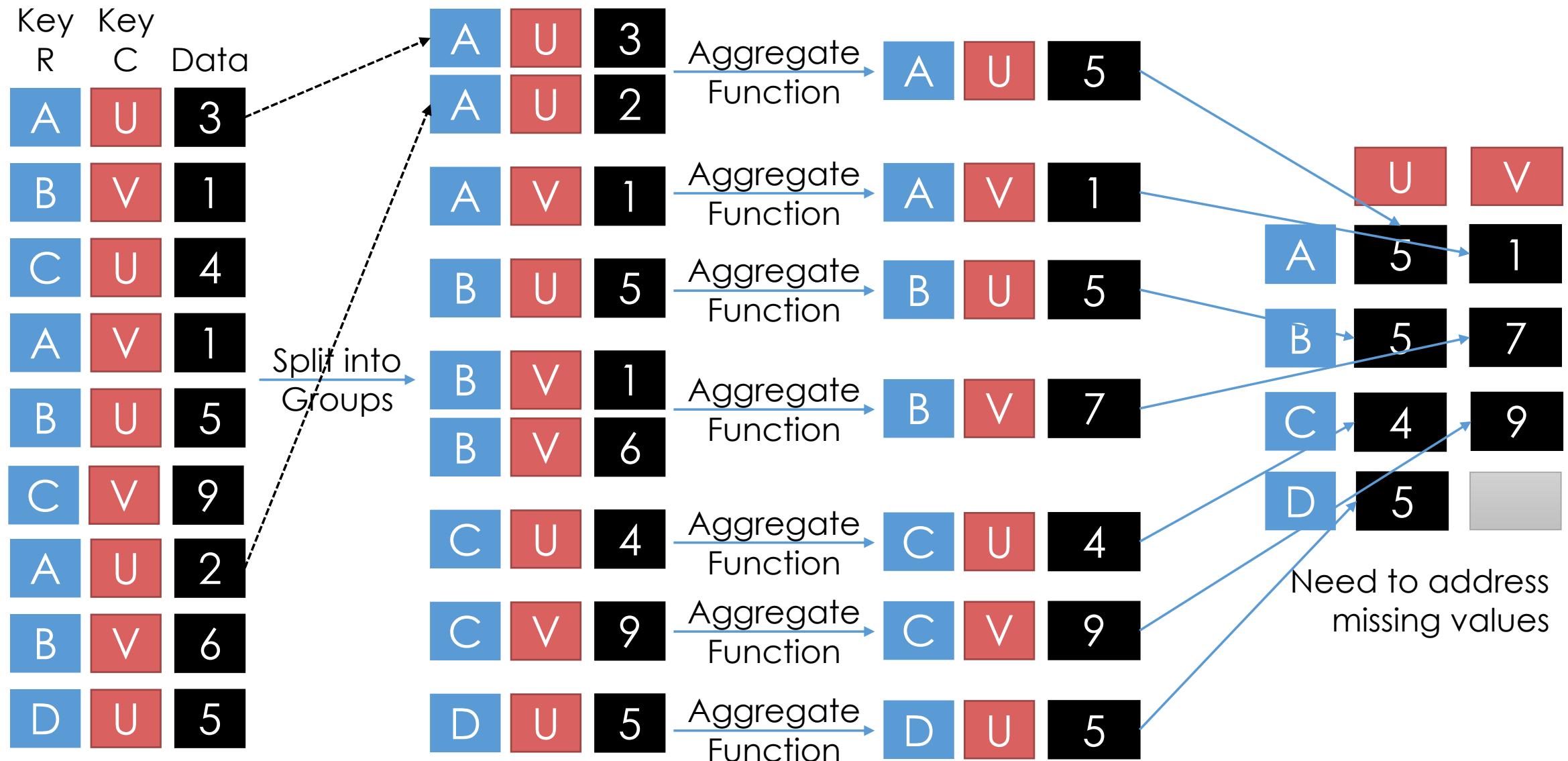
Aggregate Function

C	18
---	----

Merge Results

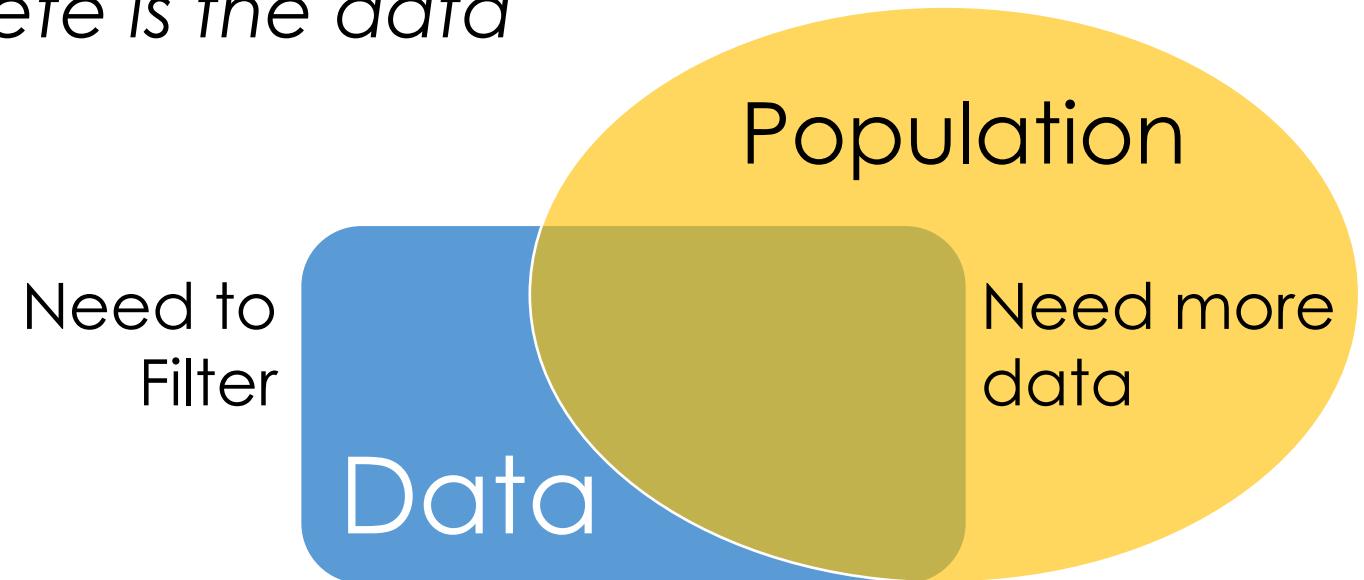
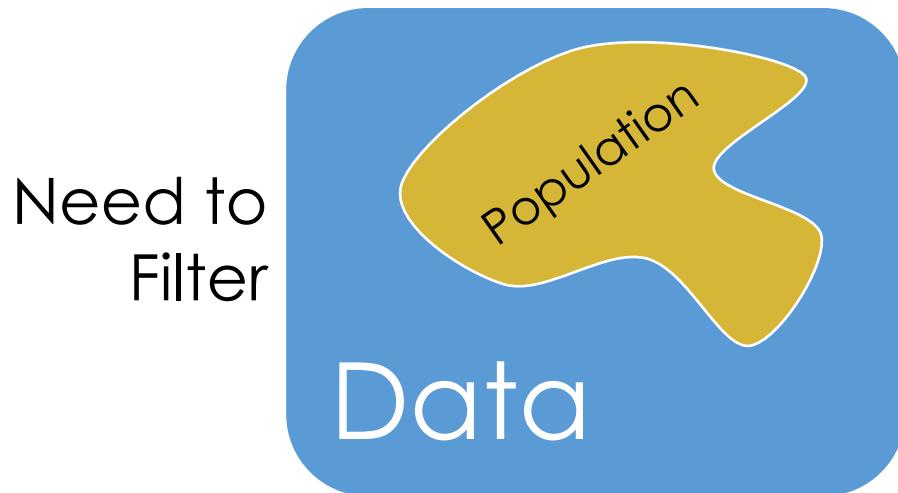
A	6
B	12
C	18

Pivot – A kind of Group By Operation



Last Lecture

- Started discussing exploratory data analysis
- **Structure** -- *the “shape” of a data file (how is it organized)*
- **Granularity** -- how fine/coarse is each datum
- **Scope** -- how (in)complete is the data



Last Lecture

- Started discussing exploratory data analysis
- **Structure** -- the “shape” of a data file (*how is it organized*)
- **Granularity** -- how fine/coarse *is each datum*
- **Scope** -- how (*in*)complete *is the data*
- **Temporality** -- *how is the data situated in time*

Temporality

- Data changes → When was the data collected!
- What is the meaning of the time and date fields?
 - When the “event” **happened**?
 - When the data was **collected** or was **entered** into the system?
 - Date the data was copied into a database (look for many matching timestamps)
- Time depends on where! (Time zones & daylight savings)
 - Learn to use **datetime** python library
 - Multiple string representation (depends on region): 07/08/09?
- Are there strange null values?
 - January 1st 1970, January 1st 1900
- Is there periodicity? Diurnal patterns

Unix Time / POSIX Time

- Time **measured in seconds** since January 1st 1970
 - Minus leap seconds ...
- Unix time follows Coordinated Universal Time (UTC)
 - International time standard
 - Measured at 0 degrees latitude
 - Similar to Greenwich Mean Time (GMT)
 - No daylight savings
 - Time codes
- Time Zones:
 - San Francisco (UTC-8)
without daylight savings



Key Data Properties to Consider in EDA

- **Structure** -- the “shape” of a data file
- **Granularity** -- how fine/coarse is each datum
- **Scope** -- how (in)complete is the data
- **Temporality** -- how is the data situated in time
- **Faithfulness** -- how well does the data capture “reality”

Key Data Properties to Consider in EDA

- **Structure** -- the “shape” of a data file
- **Granularity** -- how fine/coarse is each datum
- **Scope** -- how (in)complete is the data
- **Temporality** -- how is the data situated in time
- **Faithfulness** -- how well does the data capture “reality”

Faithfulness: Do I trust this data?

- Does my data contain unrealistic or “incorrect” values?
 - Examples?
 - Dates in the future for events in the past
 - Locations that don’t exist
 - Negative counts
 - Misspellings of names
 - Large outliers
- Does my data violate obvious dependencies?
 - E.g., age and birthday don’t match
- Was the data entered by hand?
 - Spelling errors, fields shifted ...
 - Did the form require fields or provide default values?
- Are there obvious signs of curb stoning (data falsification):
 - Repeated names, fake looking email addresses, repeated use of uncommon names or fields.

Signs that your data may not be faithful

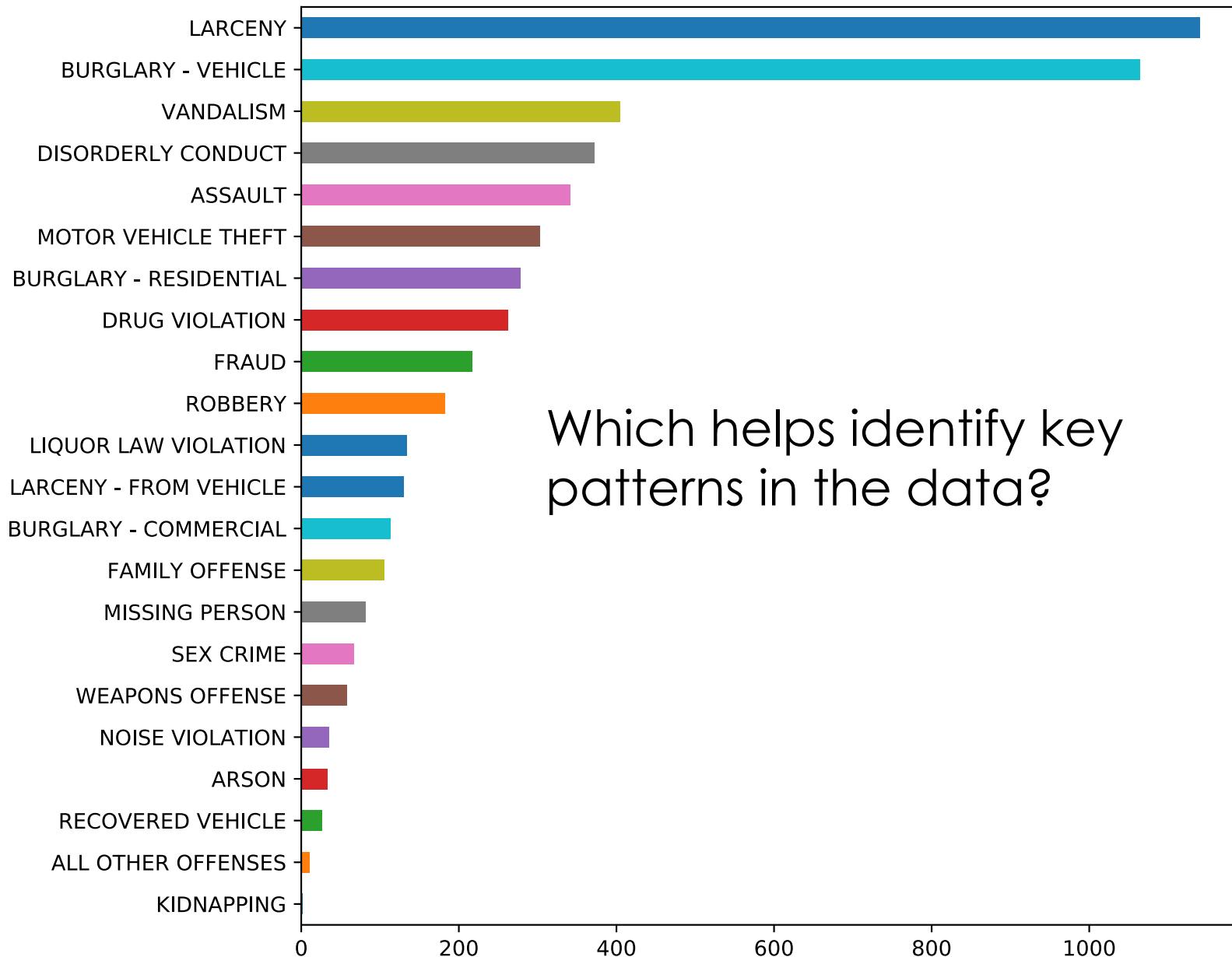
- Missing Values/Default values: (0, -1, 999, 12345, NaN, Null, 1970, 1900, ... others?)
 - **Soln 1:** Drop records with missing values → implications on your sample!
 - **Soln 2:** Impute missing values → Bias your conclusions
- Time Zone Inconsistencies
 - **Soln 1:** convert to a common timezone (e.g., UTC)
 - **Soln 2:** convert to the timezone of the location – useful in modeling behavior.
- Duplicated Records or Fields
 - **Soln:** identify and eliminate (use primary key) → implications on sample?
- Spelling Errors
 - **Soln:** Apply corrections or drop records not in a dictionary → implications on sample?
- Units not specified or consistent
 - **Solns:** Infer units, check values are in reasonable ranges for data
- Truncated data (early excel limits: 65536 Rows, 255 Columns)
 - **Soln:** be aware of consequences in analysis → how did truncation affect sample?
- Others...

How do you do EDA?

- Examine data and meta-data:
 - What is the date, size, organization, and structure of the data?
- Examine each field/attribute/dimension individually
- Examine pairs of related dimensions
 - Stratifying earlier analysis: break down grades by major ...
- Along the way:
 - Visualize/summarize the data
 - Validate assumptions about data and collection process
 - Identify and address anomalies
 - Apply data transformations and corrections
 - **Record everything you do! (why?)**

Visualization and EDA

Berkeley Crime Data



Which helps identify key patterns in the data?

ALL OTHER OFFENSES	10
ARSON	33
ASSAULT	341
BURGLARY - COMMERCIAL	113
BURGLARY - RESIDENTIAL	278
BURGLARY - VEHICLE	1064
DISORDERLY CONDUCT	372
DRUG VIOLATION	262
FAMILY OFFENSE	105
FRAUD	217
KIDNAPPING	1
LARCENY	1140
LARCENY - FROM VEHICLE	130
LIQUOR LAW VIOLATION	134
MISSING PERSON	81
MOTOR VEHICLE THEFT	303
NOISE VIOLATION	35
RECOVERED VEHICLE	26
ROBBERY	182
SEX CRIME	67
VANDALISM	404
WEAPONS OFFENSE	58
Name: CVLEGEND, dtype: int64	

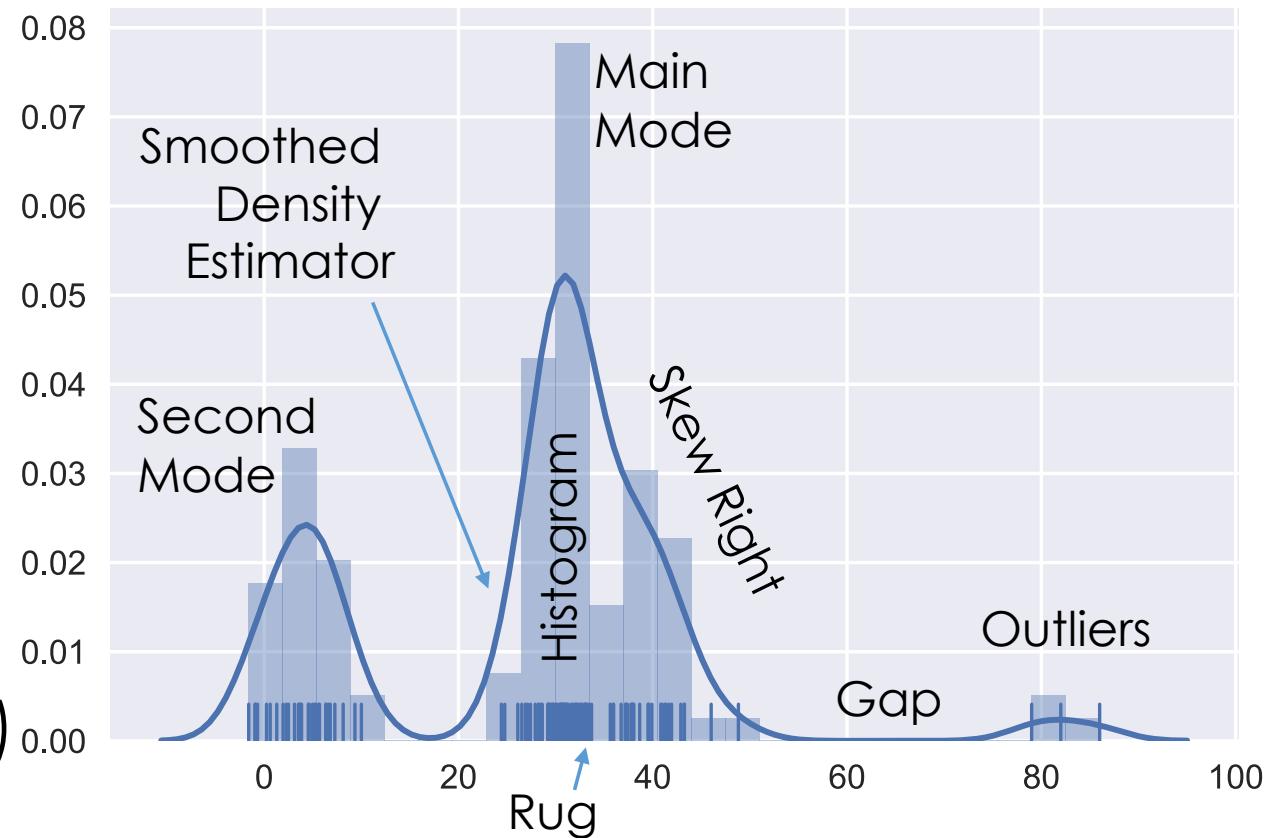
Visualizing Univariate Relationships

- **Quantitative Data**
 - Histograms, Box Plots, Rug Plots, Smoothed Interpolations (KDE – Kernel Density Estimators)
 - Look for spread, shape, modes, outliers, unreasonable values ...
- **Nominal & Ordinal Data**
 - Bar plots (sorted by frequency or ordinal dimension)
 - Look for skew, frequent and rare categories, or invalid categories
 - Consider grouping categories and repeating analysis

Histograms, Rug Plots, and KDE Interpolation

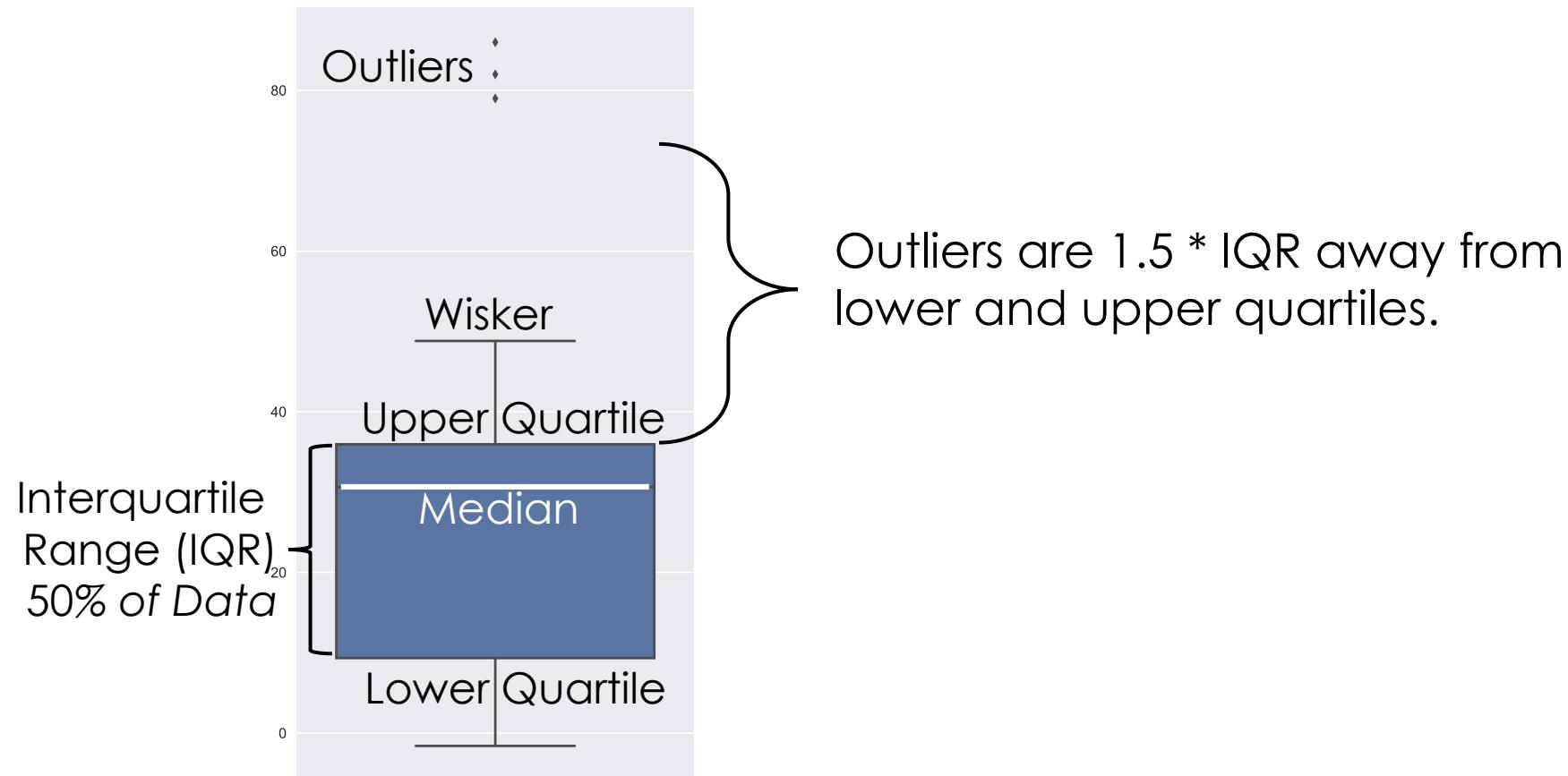
Describes distribution of data – relative prevalence of values

- Histogram
 - relative frequency of values
 - Tradeoff of bin sizes
- Rug Plot
 - Shows the actual data locations
- Smoothed density estimator
 - Tradeoff of “bandwidth” parameter (more on this later)



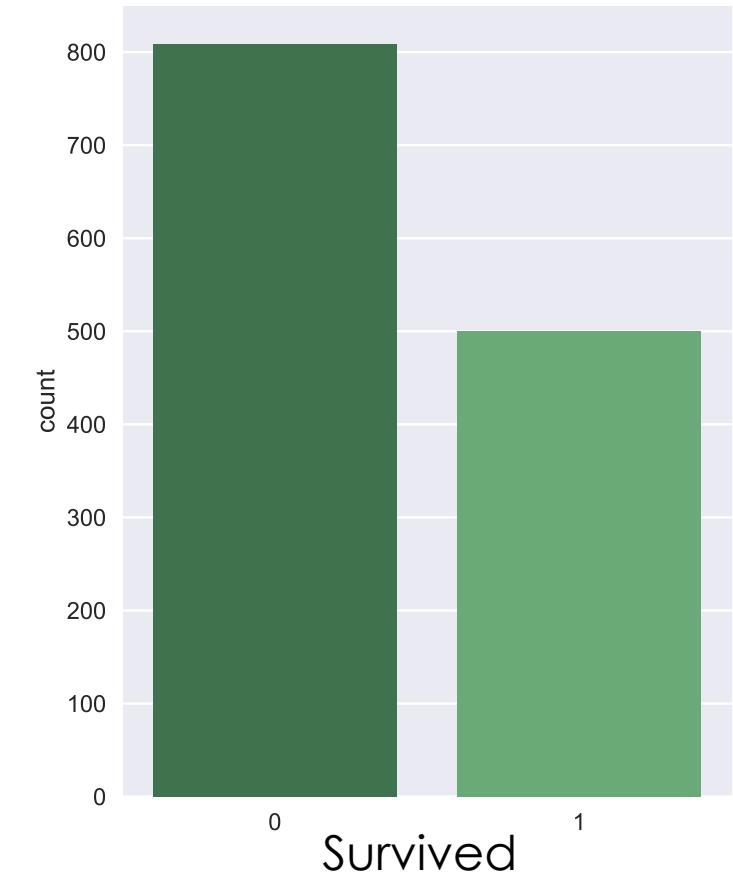
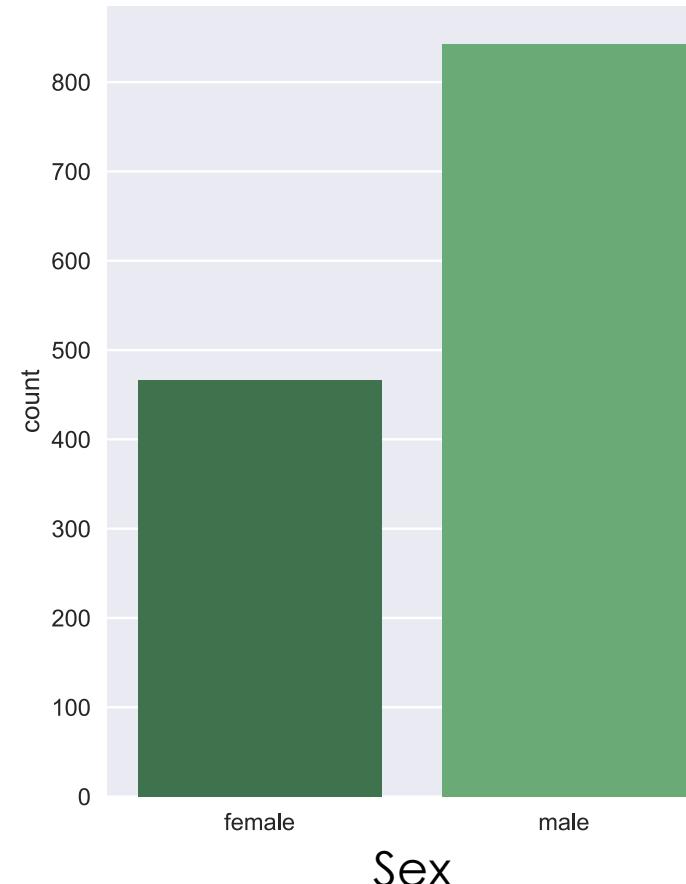
Box Charts

- Useful for summarizing distributions and comparing multiple distributions



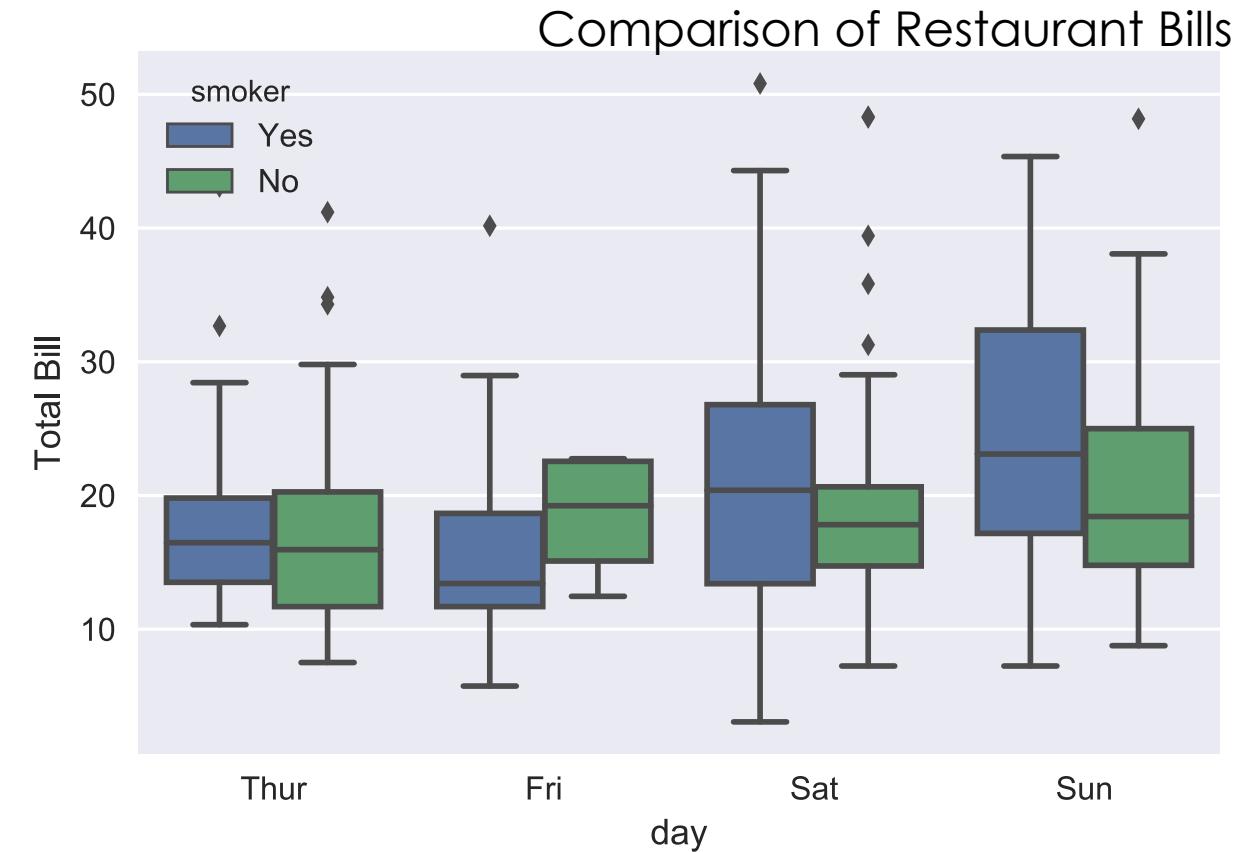
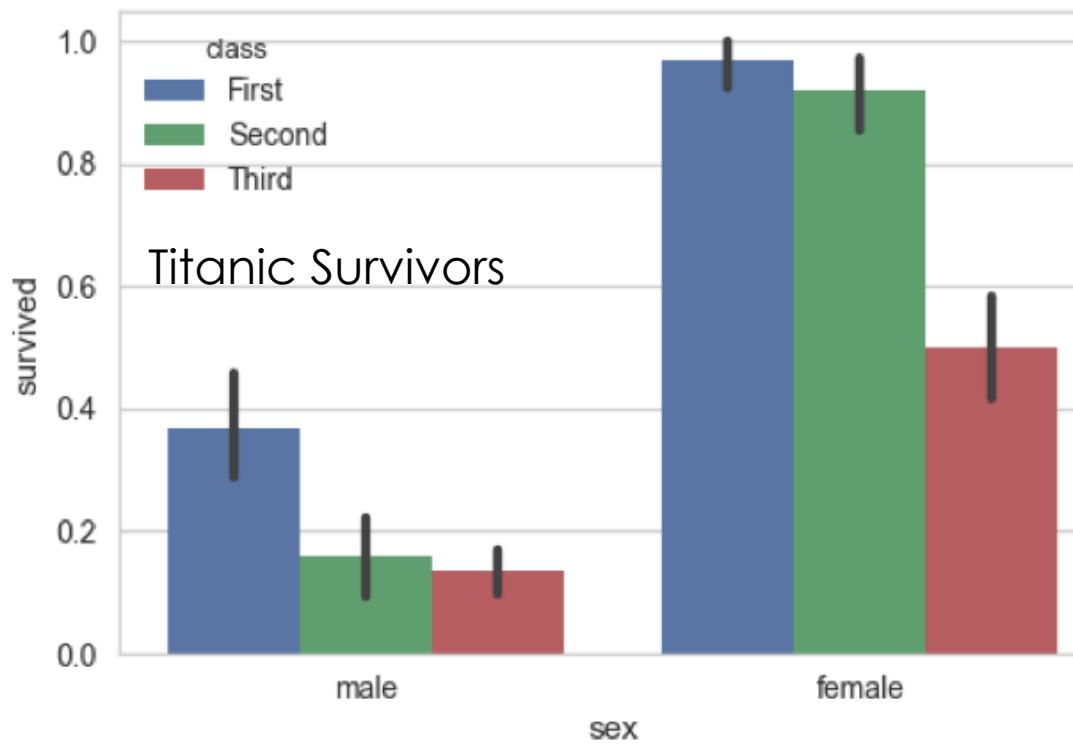
Bar Charts

- Used to compare nominal and ordinal data.
- Consider sorting by category or frequency

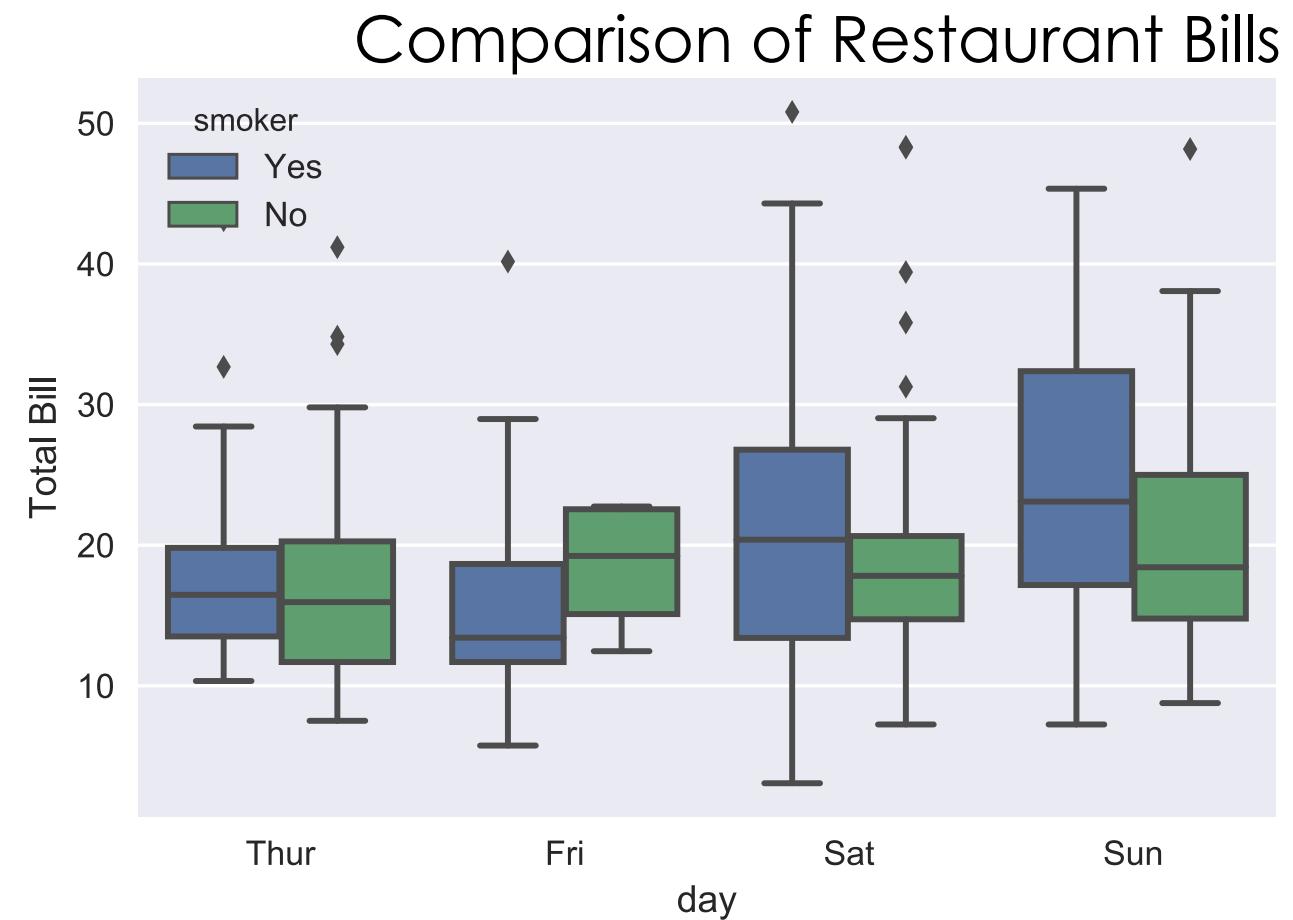
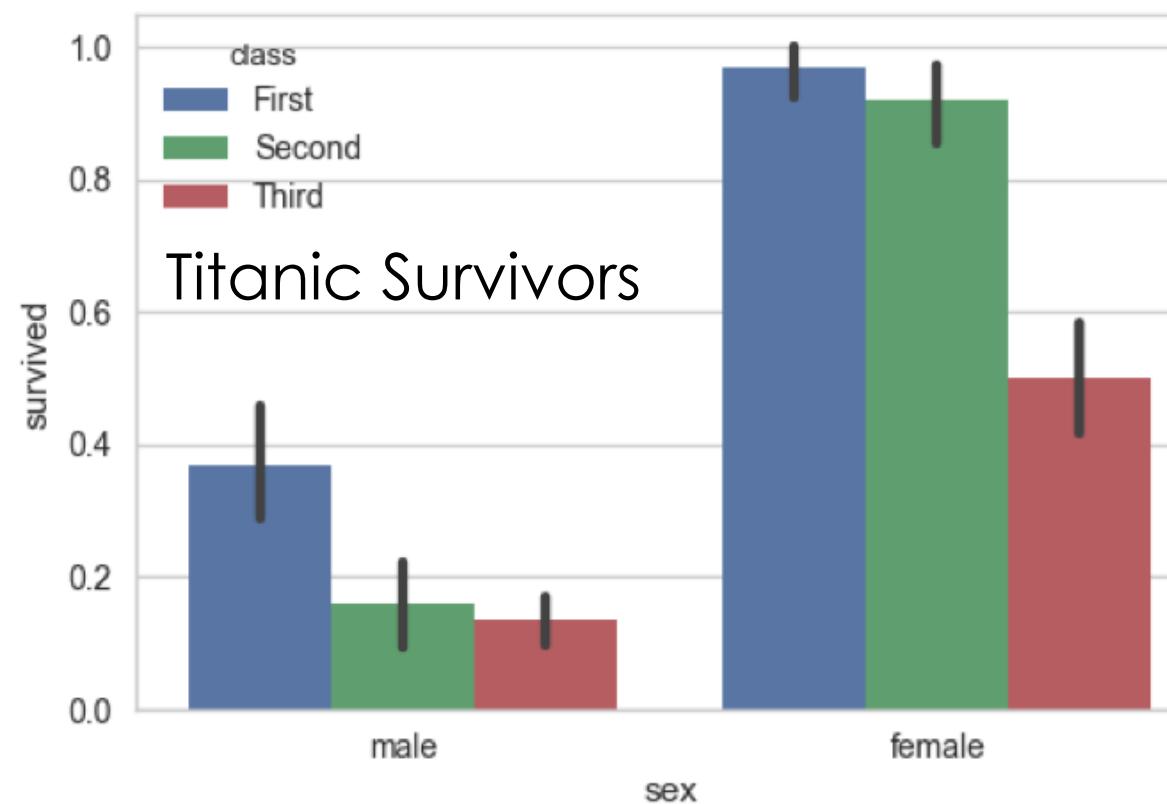


Visualizing Multivariate Relationships

- Conditioning on a range of values (e.g., ages in groups) and construct side by side box-plots or bar charts



<http://bit.ly/ds100-sp18-eda>

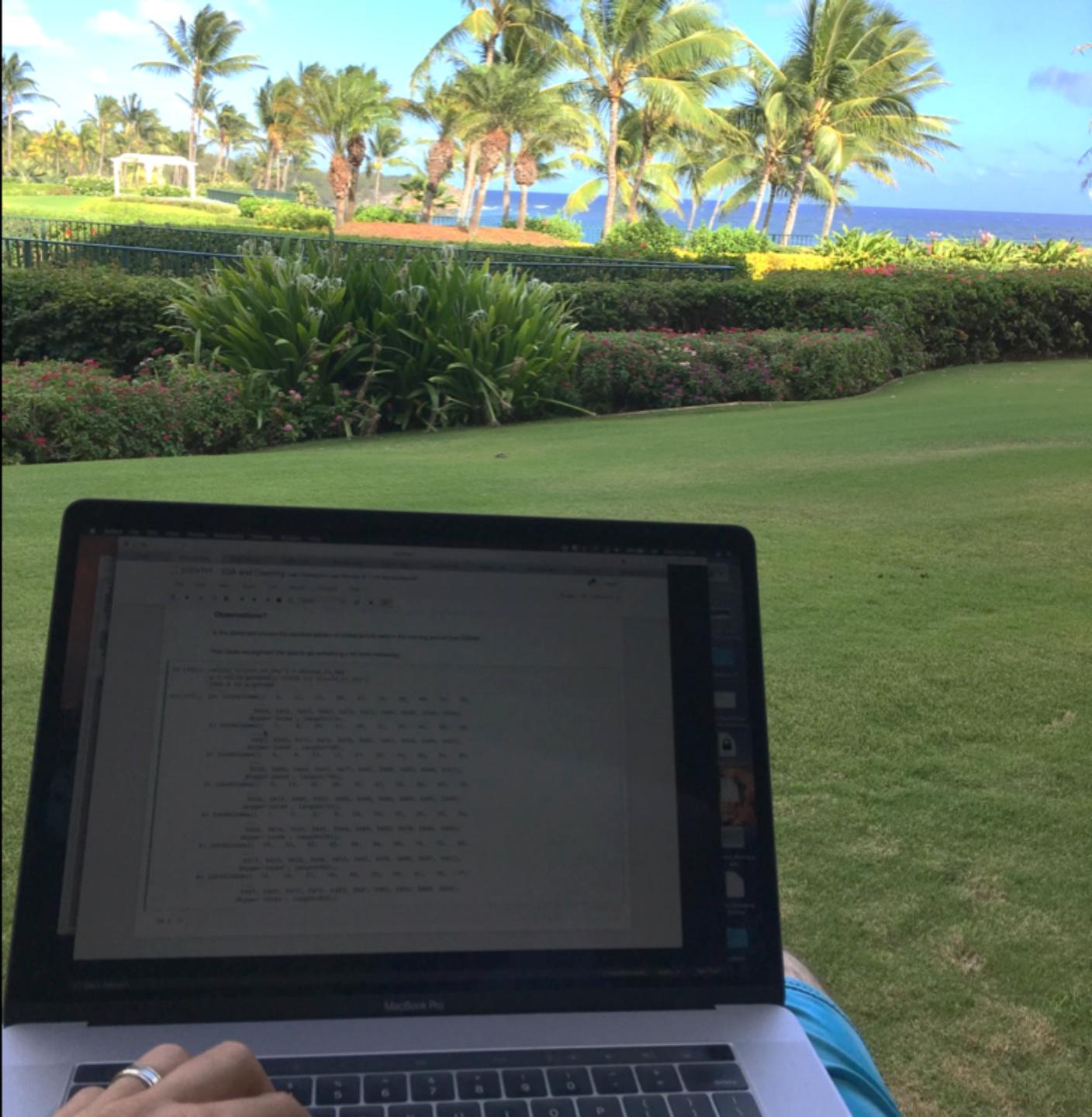


Quick Break



Quick Break

Scope:
Do you have a full picture?



Berkeley Police Data Demo

Berkeley Police Public Datasets

- **Question:** For this analysis we will not begin with a detailed question but instead a rough goal of understanding Police activity.
- **Examine Two Data Sets:**
 - Call data
 - Stop data
- Today we will work through the basic process of data loading, some preliminary cleaning, and exploratory data analysis.

Call Data Description

Data pulled from Public Safety Server using data created for Berkeley's Crime View Community page. Displays **incidents reported** for **the last 180 days** along with **time, date, day of week** and **block level location information**.

The dataset reflects crimes as they have been reported to the BPD based on preliminary information **supplied by the reporting parties**. Preliminary crime classifications may change based on follow-up investigations. **Not all calls for police service are included (e.g. Animal Bite)**. The information provided on this site is intended for use by the community to enhance their awareness of crimes occurring in their neighborhoods and the entire City. **The data should not be used for in-depth crime analysis** as the initial information is subject to change.

Stops Data Description

This data was extracted from the Department's Public Safety Server and covers the **data beginning January 26, 2015**. On January 26, 2015 the department began collecting data pursuant to General Order B-4 (issued December 31, 2014). Under that order, **officers were required to provide certain data after making all vehicle detentions** (including bicycles) and pedestrian detentions (up to five persons). This data **set lists stops by police** in the categories of traffic, suspicious vehicle, pedestrian and bicycle stops. Incident number, date and time, location and disposition codes are also listed in this data.

Address **data has been changed from a specific address**, where applicable, and listed as the block where the incident occurred. Disposition codes were entered by officers who made the stop. These codes included the person(s) race, gender, age (range), reason for the stop, enforcement action taken, and whether or not a search was conducted.

Caution about EDA

With enough data, if you look hard enough you will find something “**interesting**”

Important to differentiate **inferential conclusions** about world from **exploratory analysis of data**

