SF-HIP Statistical Analysis

Kris, Terence, Chao, Ray, and Violet

March 29, 2015

```
## [[1]]
## [1] TRUE
##
## [[2]]
## [1] TRUE
##
## [[3]]
## [1] TRUE
##
## [[4]]
## [1] TRUE
##
## [[5]]
## [1] TRUE
##
## [[6]]
## [1] TRUE
##
## [[7]]
## [1] TRUE
##
## [[8]]
## [1] TRUE
```

Introduction

In this document, we collect the statistical approaches and insights from DataKind's March 27-29, 2015 DataDive. While we highlight our most interesting findings, we also wanted to share alternative lines of thought we had developed but not completely refined, with the hope that this could guide future analysis.

Goals

There are two overall goals of interest in this analysis: Global patterns and local anomalies. On the one hand, we would like to summarise the relationships between the presence of liquor stores and crimes that will apply generally across San Francisco. On the other hand, we would like to highlight those locations which somehow seem to deviate from any general patterns. In both cases, we would like to integrate demographic information. For example, are there locations with similar demograhpics and numbers of liquor stores, but very different crime rates?

Approach

Data Available

We consider three primary sources of data: - Census information: Demographic information at the census tract level. This includes overall population, population breakdown across races, unemployment rate, and

median income. - Crime data: We have crime reports (from the SFPD?), mapping the time and place of crimes within the city over the last 10 years. These crime reports also include descriptions of the type of crime at varying levels of granularity – a report may be classified at a coarse level as robbery, and at a fine level as robbery at an ATM machine, for instance. - Alcohol license data: We have records of alcohol licenses over more than a decade. These licenses are required by any venue that sells or distributes alcohol, including bars, clubs, and convenience stores. These records include the location of these vendors, as well as a license type (bars and liquor stores require different licenses, for example).

Data used

We chose to aggregate the crime and alcohol license data to the census tract level, and then normalize by census tract population. More specifically, we (1) counted the number of venues using each of the 23 license types within each census tract, then divided by the population of that census tract and (2) counted the number of crimes within each of the 30 description groups. We could have used finer or coarser description types for both liquor vendors and crimes, but this level seemed to offer a rich description without making the problem too high dimensional, and less tractable. Further, we discarded those tracts with fewer than 500 people living within them, since our estimates of the densities of crimes and liquor venues in such sparsely populated areas are less reliable.

Notice that, at this stage, we have ignored (1) any spatial information at a finer resolution than the census tract level and (2) any temporal effects. Nonetheless, we believe our methods could be generalized to handle these situations as well.

Methods

For the first task, identifying global patterns, our overal methods use two steps: dimension reduction followed by some measure of association. By dimension reduction, we mean reducing many different measurements to just a few – for example, the number of college educated people and the median income of a census tract can both be explained by an underlying "affluence" effect. By association, we mean taking these underlying factors and determining whether and how they are correlated.

The specific tools we applied to do this reduction vary in complexity. From crudest to most (but still not very) sophisticated, we used

-Crudest dimension reduction is just summing counts - Another crude dimension reduction is ignoring dimensions - Once we cluster based on two sets of vars, see how the clusterings compare - See how the distances compare - Perform dimension reduction jointly

Regression on Aggregated Counts

- First, just summing counts of licenses
- We need to compare rates
- We find that most rates are very small, with a few getting much larger, so we use a log scale
- As a response, look at the liquor and all crimes
- Make plots, just to see association, follow-up with formal regressions to validate visual intuition
- Regress with control covariates

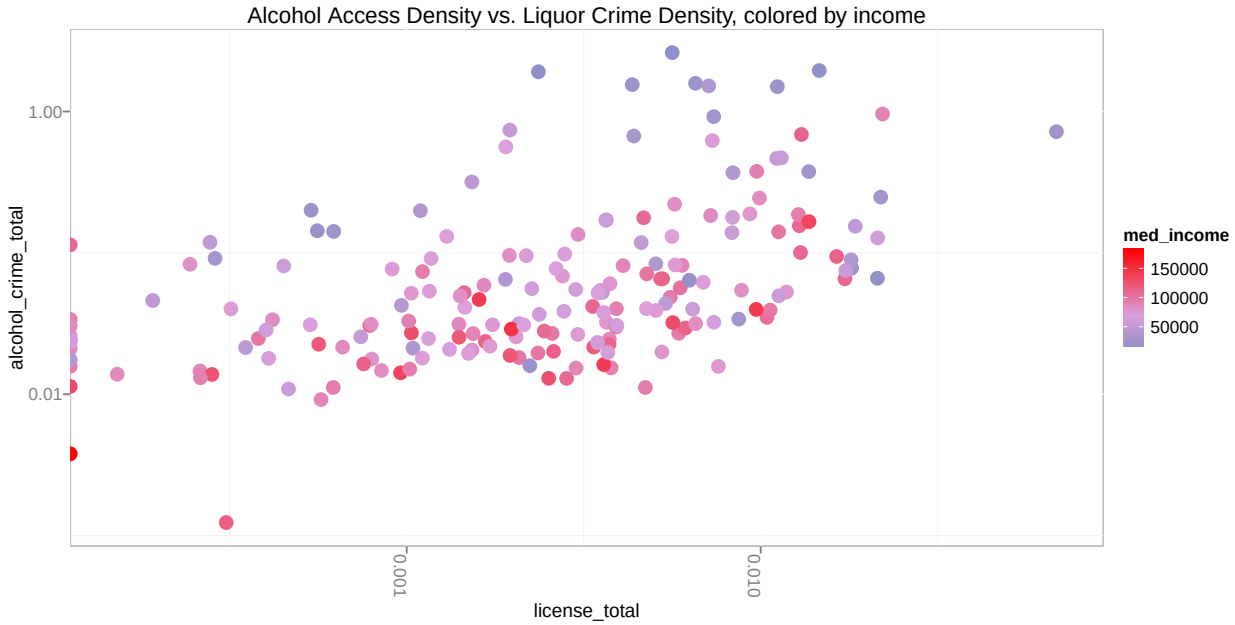
Liquor Crimes —

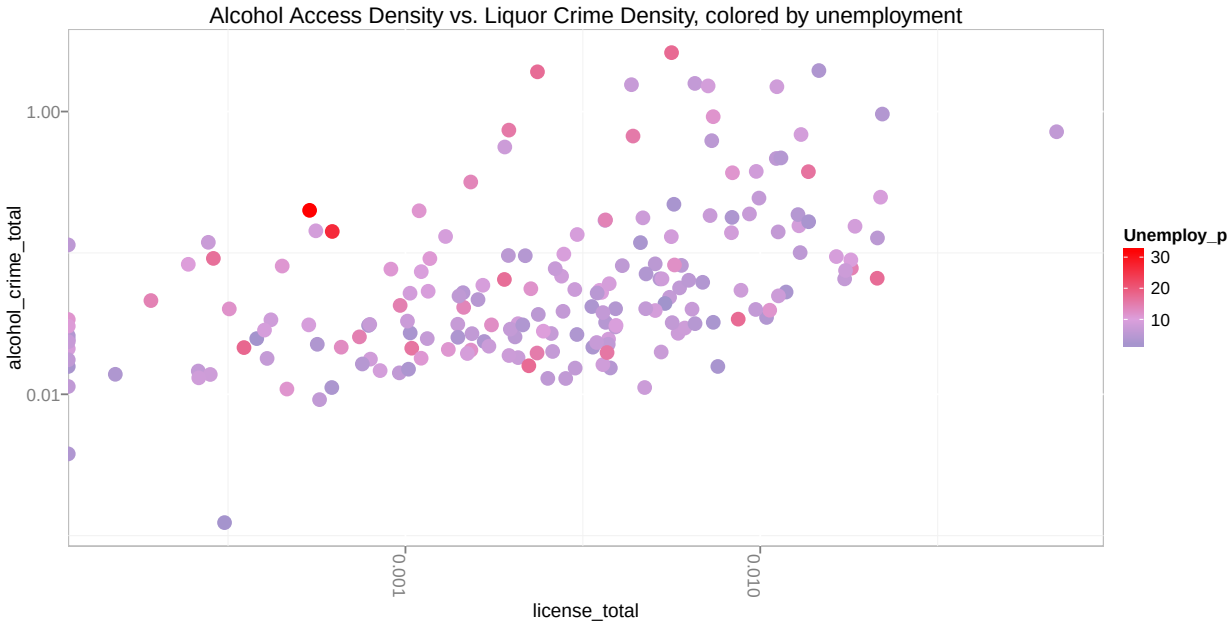
- Code for liquor crimes only

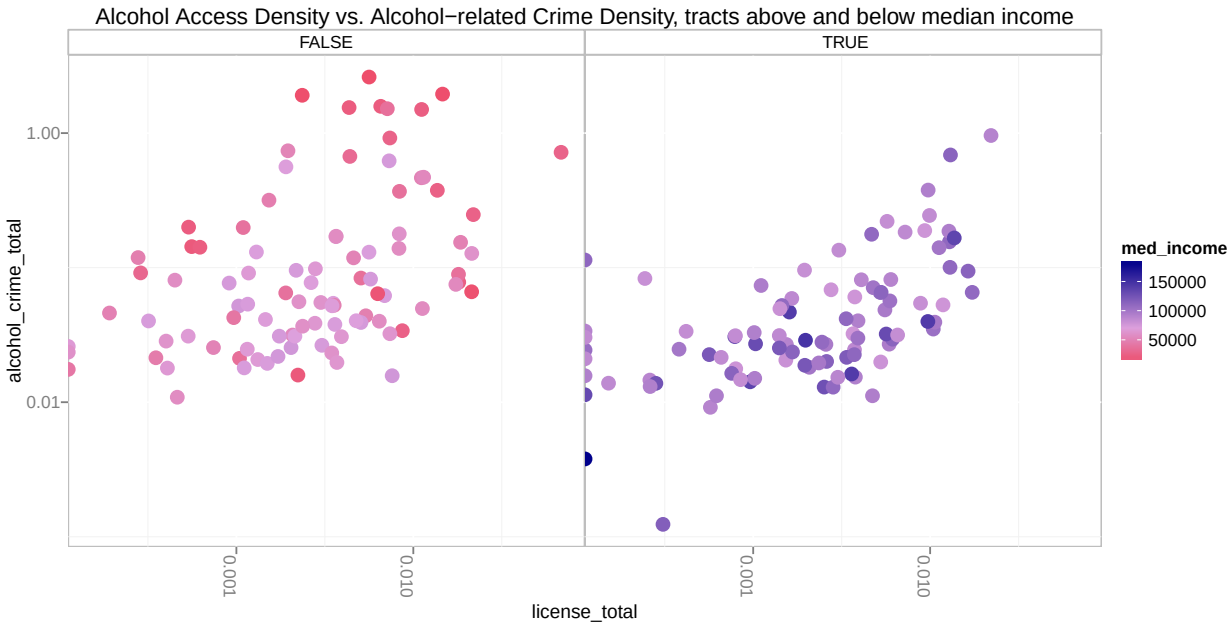
Setup, since haven't run code yet.

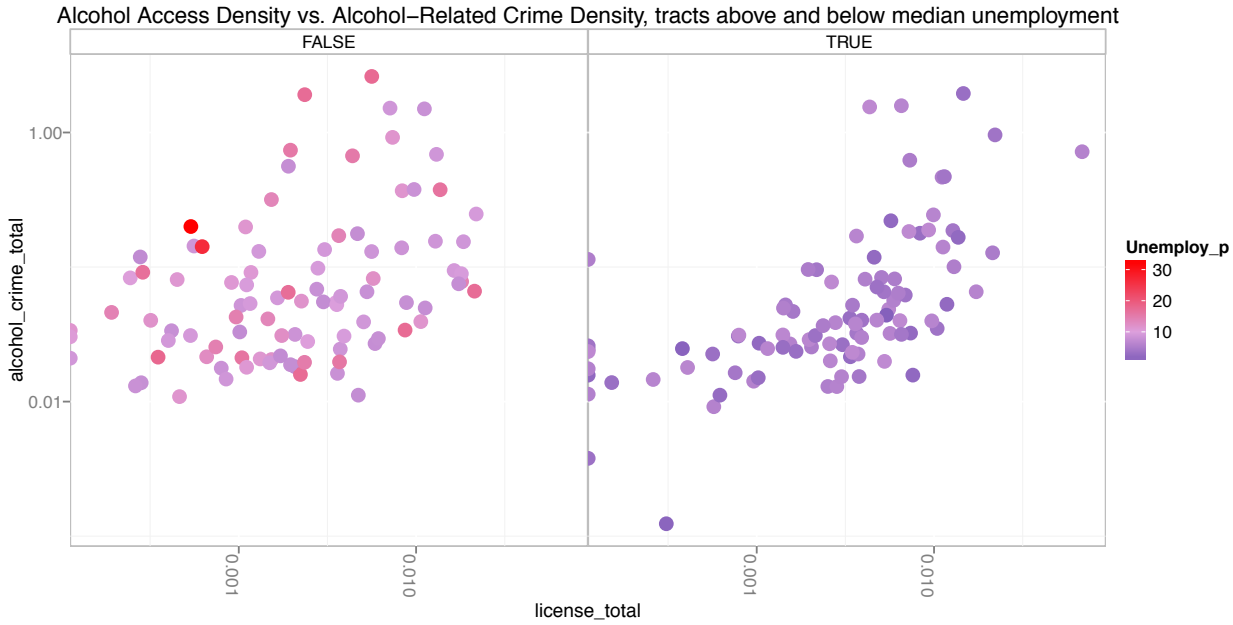
Alcohol Related Crimes —

- Interpretation of plot, for all grouped together
- Increasing density of liquor establishments is associated with an increase in number of liquor related crimes (mention what they are)
- For a fixed density of liquor establishments, there is still a substantial variation across crime rates
- Condition on income level, unemployment
- Look at shaded plots: pattern is not so strong by eye. Different colors mean different levels of income, unemployment within census tract
- Look at faceted plots
- Interpretations
- Seems like the effect size of increases in liquor store density on liquor crime density are larger in neighorhoods with below median income
- Run regressions to quantify these differences
- Run joint regression y ~ beta0 + (beta1 + beta2 * I(above median)) * liquor
  - This turns out to not be significant. So, even though visually we see one pattern, this is not formally statistically significant
- Run joint regression y ~ beta0 + beta1 * liquor + beta2 * median income
  - Idea is now can interpret liquor store effect controlling for median income




Alcohol Access Density vs. Liquor Crime Density, colored by unemployment
alcohol_crime_total
1.00
0.01
0.001
0.010
license_total
Unemploy_p
30
20
10


Alcohol Access Density vs. Alcohol−related Crime Density, tracts above and below median income
FALSE
TRUE
alcohol_crime_total
1.00
0.01
0.001
0.010
0.001
0.010
license_total
med_income
150000
100000
50000




```
##
## Call:
## lm(formula = log(1 + alcohol_crime_total) ~ log(1 + license_total) +
##     log(1 + license_total) * above_median_inc, data = liquor_laws_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.24425 -0.10968 -0.03025  0.01018  1.10402
##
## Coefficients:
##                                            Estimate Std. Error t value
## (Intercept)                                 0.13307    0.02437   5.461
## log(1 + license_total)                      8.29156    2.48683   3.334
## above_median_incTRUE                       -0.12793    0.03688  -3.469
## log(1 + license_total):above_median_incTRUE  5.61279    5.16531   1.087
##                                            Pr(>|t|)
## (Intercept)                                 1.5e-07 ***
## log(1 + license_total)                      0.00103 **
## above_median_incTRUE                        0.00065 ***
## log(1 + license_total):above_median_incTRUE  0.27861
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1975 on 186 degrees of freedom
## Multiple R-squared:  0.1648, Adjusted R-squared:  0.1513
## F-statistic: 12.23 on 3 and 186 DF,  p-value: 2.431e-07

##
## Call:
## lm(formula = log(1 + alcohol_crime_total) ~ log(1 + license_total) +
##     log(1 + license_total) * below_median_unemp, data = liquor_laws_data)
##
## Residuals:
```

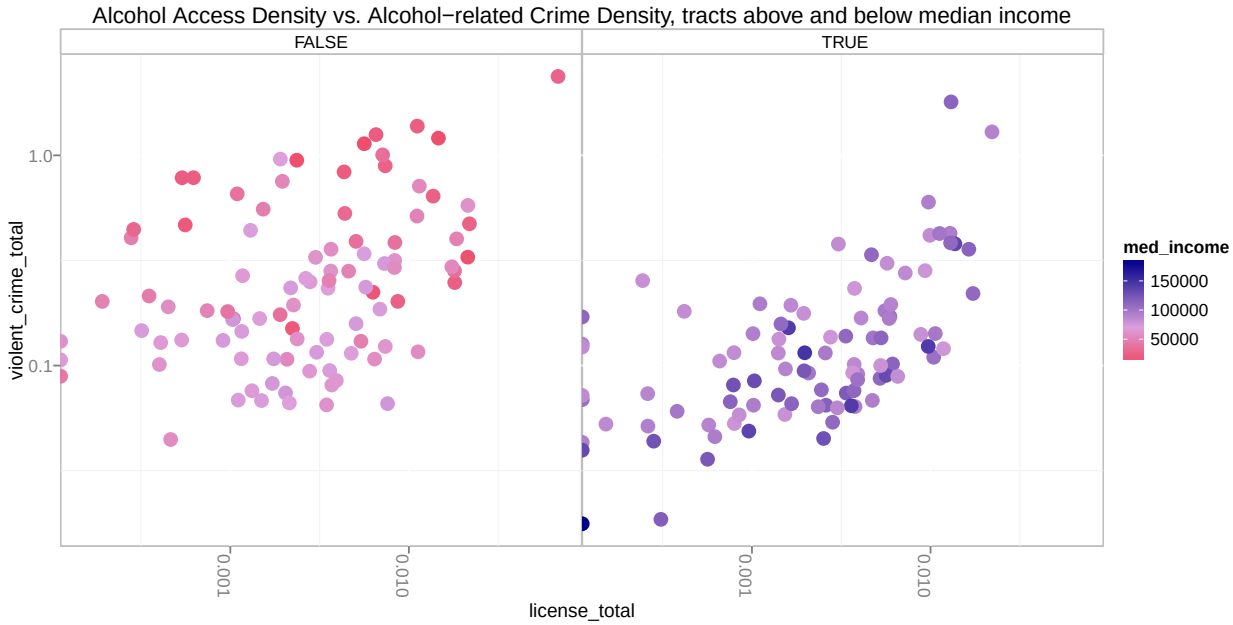
```
##      Min       1Q   Median       3Q      Max
## -0.25061 -0.08247 -0.04883 -0.02053  1.13529
##
## Coefficients:
##                                             Estimate Std. Error t value
## (Intercept)                                  0.09809    0.02723   3.602
## log(1 + license_total)                       8.95202    4.03642   2.218
## below_median_unempTRUE                      -0.05333    0.03715  -1.436
## log(1 + license_total):below_median_unempTRUE  2.34843    4.85346   0.484
##                                             Pr(>|t|)
## (Intercept)                                 0.000405 ***
## log(1 + license_total)                      0.027778 *
## below_median_unempTRUE                      0.152816
## log(1 + license_total):below_median_unempTRUE 0.629050
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2035 on 186 degrees of freedom
## Multiple R-squared:  0.1131, Adjusted R-squared:  0.09875
## F-statistic: 7.903 on 3 and 186 DF,  p-value: 5.454e-05

##
## Call:
## lm(formula = log(1 + alcohol_crime_total) ~ log(1 + license_total) +
##     log(1 + license_total) + med_income, data = liquor_laws_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.35336 -0.08848 -0.03560  0.02469  0.98187
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             2.889e-01  3.937e-02   7.339 6.38e-12 ***
## log(1 + license_total)  7.852e+00  2.088e+00   3.760 0.000227 ***
## med_income             -2.684e-06  4.387e-07  -6.117 5.47e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1864 on 187 degrees of freedom
## Multiple R-squared:  0.2518, Adjusted R-squared:  0.2438
## F-statistic: 31.47 on 2 and 187 DF,  p-value: 1.66e-12

##
## Call:
## lm(formula = log(1 + alcohol_crime_total) ~ log(1 + license_total) +
##     Unemploy_p, data = liquor_laws_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.30947 -0.08124 -0.04335 -0.01403  1.07486
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             0.003294   0.032271   0.102   0.9188
```

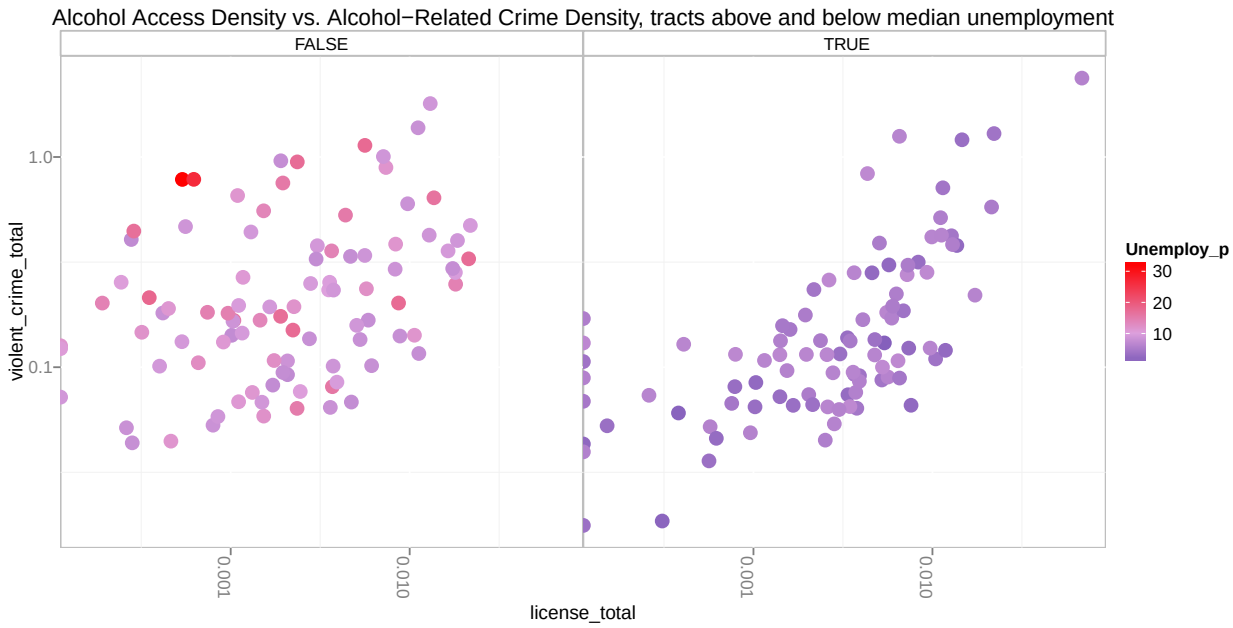
```
## log(1 + license_total) 10.758539   2.212033   4.864 2.44e-06 ***
## Unemploy_p              0.008296   0.003284   2.526   0.0124 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2008 on 187 degrees of freedom
## Multiple R-squared:  0.1318, Adjusted R-squared:  0.1225
## F-statistic: 14.19 on 2 and 187 DF,  p-value: 1.833e-06
```

```
##
## Call:
## lm(formula = log(1 + alcohol_crime_total) ~ log(1 + license_total) +
##     Unemploy_p + log(1 + license_total) * Unemploy_p, data = liquor_laws_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.28353 -0.08326 -0.04347 -0.00810  1.09242
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     -0.017090   0.036276  -0.471   0.6381
## log(1 + license_total)          16.274463   5.018673   3.243   0.0014 **
## Unemploy_p                       0.011031   0.003968   2.780   0.0060 **
## log(1 + license_total):Unemploy_p -0.774204   0.632499  -1.224   0.2225
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2005 on 186 degrees of freedom
## Multiple R-squared:  0.1387, Adjusted R-squared:  0.1248
## F-statistic: 9.983 on 3 and 186 DF,  p-value: 3.917e-06
```

Violent Crimes—

- What about violent crimes? Repeat same bullets as above






```
##
## Call:
## lm(formula = log(1 + violent_crime_total) ~ log(1 + license_total) +
##     log(1 + license_total) * above_median_inc, data = liquor_laws_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.24339 -0.08895 -0.02543  0.03807  0.68436
##
## Coefficients:
##                                          Estimate Std. Error t value
## (Intercept)                               0.19862    0.01918  10.355
## log(1 + license_total)                   13.82449    1.95764   7.062
## above_median_incTRUE                     -0.14712    0.02903  -5.068
```

```
## log(1 + license_total):above_median_incTRUE  8.74409    4.06613   2.150
##                                             Pr(>|t|)
## (Intercept)                                  < 2e-16 ***
## log(1 + license_total)                      3.18e-11 ***
## above_median_incTRUE                        9.68e-07 ***
## log(1 + license_total):above_median_incTRUE   0.0328 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1554 on 186 degrees of freedom
## Multiple R-squared:  0.3968, Adjusted R-squared:  0.3871
## F-statistic: 40.78 on 3 and 186 DF,  p-value: < 2.2e-16

##
## Call:
## lm(formula = log(1 + violent_crime_total) ~ log(1 + license_total) +
##     log(1 + license_total) * below_median_unemp, data = liquor_laws_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.21351 -0.09038 -0.03337  0.03173  0.68389
##
## Coefficients:
##                                             Estimate Std. Error t value
## (Intercept)                                  0.19157    0.02135   8.971
## log(1 + license_total)                      11.78494    3.16485   3.724
## below_median_unempTRUE                      -0.11745    0.02913  -4.032
## log(1 + license_total):below_median_unempTRUE  7.66712    3.80547   2.015
##                                             Pr(>|t|)
## (Intercept)                                 3.15e-16 ***
## log(1 + license_total)                       0.00026 ***
## below_median_unempTRUE                      8.05e-05 ***
## log(1 + license_total):below_median_unempTRUE  0.04537 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1596 on 186 degrees of freedom
## Multiple R-squared:  0.3645, Adjusted R-squared:  0.3543
## F-statistic: 35.56 on 3 and 186 DF,  p-value: < 2.2e-16

##
## Call:
## lm(formula = log(1 + violent_crime_total) ~ log(1 + license_total) +
##     log(1 + license_total) + med_income, data = liquor_laws_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.31674 -0.08308 -0.02565  0.03610  0.79126
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             3.323e-01  3.102e-02  10.714  < 2e-16 ***
## log(1 + license_total)  1.432e+01  1.645e+00   8.707 1.62e-15 ***
## med_income             -2.494e-06  3.457e-07  -7.214 1.31e-11 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1469 on 187 degrees of freedom
## Multiple R-squared:  0.4587, Adjusted R-squared:  0.4529
## F-statistic: 79.22 on 2 and 187 DF,  p-value: < 2.2e-16

##
## Call:
## lm(formula = log(1 + violent_crime_total) ~ log(1 + license_total) +
##     Unemploy_p, data = liquor_laws_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33289 -0.08508 -0.02801  0.03228  0.66788
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             0.021007   0.024906   0.843      0.4
## log(1 + license_total) 17.306281   1.707233  10.137  < 2e-16 ***
## Unemploy_p              0.013336   0.002535   5.262 3.88e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.155 on 187 degrees of freedom
## Multiple R-squared:  0.3972, Adjusted R-squared:  0.3908
## F-statistic: 61.62 on 2 and 187 DF,  p-value: < 2.2e-16

##
## Call:
## lm(formula = log(1 + violent_crime_total) ~ log(1 + license_total) +
##     Unemploy_p + log(1 + license_total) * Unemploy_p, data = liquor_laws_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.21930 -0.08655 -0.02769  0.03654  0.69113
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     -0.018908   0.027378  -0.691  0.49067
## log(1 + license_total)          28.107322   3.787607   7.421 4.03e-12
## Unemploy_p                       0.018691   0.002995   6.241 2.87e-09
## log(1 + license_total):Unemploy_p -1.516013   0.477349  -3.176  0.00175
##
## (Intercept)
## log(1 + license_total)          ***
## Unemploy_p                      ***
## log(1 + license_total):Unemploy_p **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1513 on 186 degrees of freedom
## Multiple R-squared:  0.4282, Adjusted R-squared:  0.419
## F-statistic: 46.44 on 3 and 186 DF,  p-value: < 2.2e-16
```

General Patterns in Unaggregated Counts —

- Everything so far has been one dimensional, really ought to consider different crimes and liquor types separately, do more intelligent aggregation
- Some groups of license types / crime types are quantitatively grouped together [still focus on liquor related crimes]
- Further, which license types are associated with which types of crimes? Which groups of license types are associated with which groups of crimes?
- Do any tracts show an especially high level of (1) one group of licenses, or (2) one group of crimes? What about groups of tracts?
- The point is that it's easy to speak about single crime and license types at the single tract level, as well as the sum across crimes, licenses and tracts. But, we would appreciate intermediate levels of resolution if there is some grouping on these variables (but the regime is not entirely homogeneous)
- Two very general statistical tools available for this sort of reduction are multivariate analysis and clustering.

Cluster Analysis

Clustering tracts

- Define a distance between tracts, based on different kinds of data
- Which tracts group together? How similar are they across the different metrics?
- Maybe clustering tracts using similarity across sums within clustered columns?

Multivariate Analysis on Unaggregated —

- Clustering assigns a hard label to each tract
- Multivariate analysis attempts to find a lower dimensional representation of the data that doesn't lose too much information. What are the sourecs of maximum variation?

Bivarite Correlations —

- Before full multivariate analysis, consider the largest bivariate correlations
- All crimes, and just crimes of interest
- Interpretation: The highest correlation is within data groups
- But nonetheless very strong association across groups
- Consider list of top correlations across crime, license, and demographics
- List is of limited utility: It's unecessarily verbose of multiple crimes are all correlated with each other, they will all give high correlations with license types.

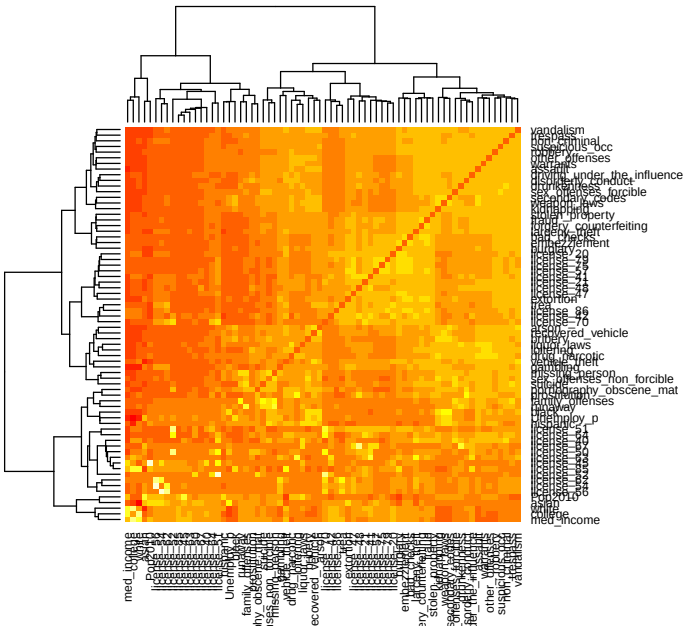
```
##              Var1                   Var2     value
## 1           fraud forgery_counterfeiting 0.9719785
## 3        warrants         other_offenses 0.9567610
## 5  other_offenses                assault 0.9526326
## 7      license_56            license_54 0.9524796
## 9    embezzlement            bad_checks 0.9501824
## 11       warrants                assault 0.9419331
```

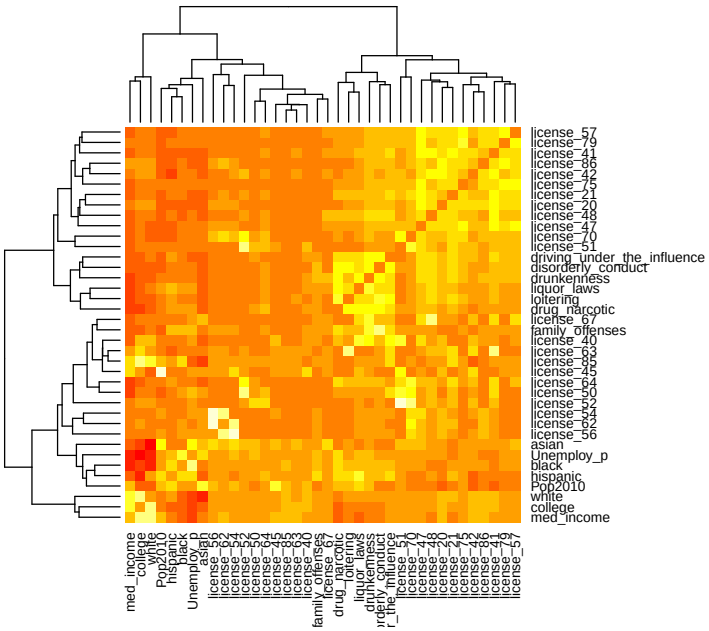
```
##                     Var1        Var2     value
## 1             license_56  license_54 0.9524796
```

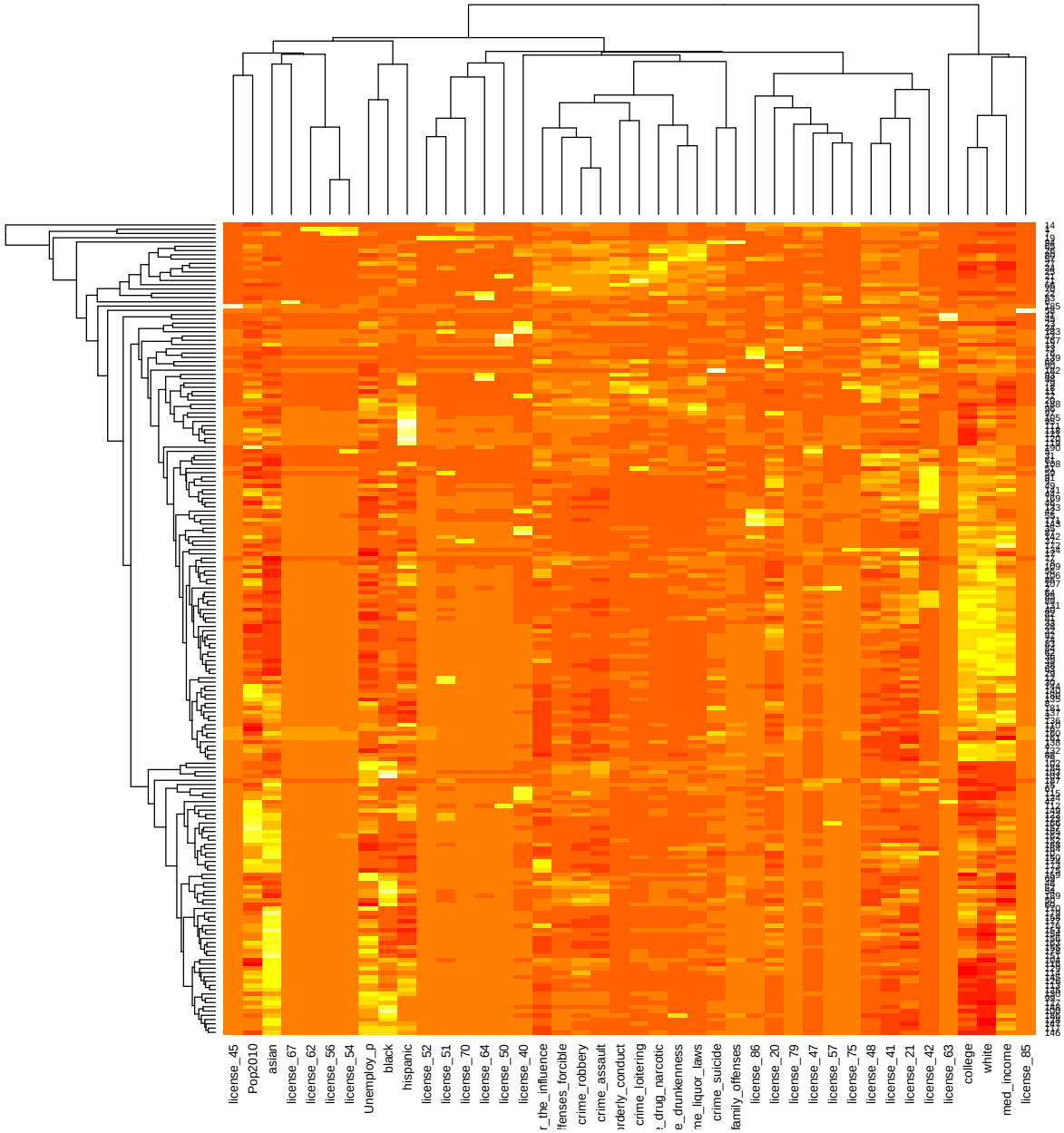
```
## 3                robbery     assault 0.9150940
## 5             license_62  license_56 0.8303807
## 7  sex_offenses_forcible     assault 0.8289806
## 9                college       white 0.8239504
## 11           drunkenness liquor_laws 0.8145121
```

Basic Clustering & Biclustering —

- A brief digression to show the biclustering of data according to both raw data and the correlations
- The interpretation is that we try to hierarchically cluster features, based on raw data or correlations between them
- Groups of correlations or groups of actual data that are most similar to each other are merged first
- This is usually just a nice preliminary visualization, not too much insight either
- Yellow is higher correlated than red
- Blocking is interesting: These are variables that can be essentially collapsed.
- But we see this pattern of crimes with each other and licenses with each other.
- Notice it's symmetric, trees are identical on left and top.
- Also, some substantial correlation across groups of cols, as expected from before
- Mention the top fiew: Median income and college. Median Income and License 85.
- 85 and 45 are merged first: These are most similar license types
- Much more interpretation is possible
- Can do the same with tracts as one of the heatmap dimensions.

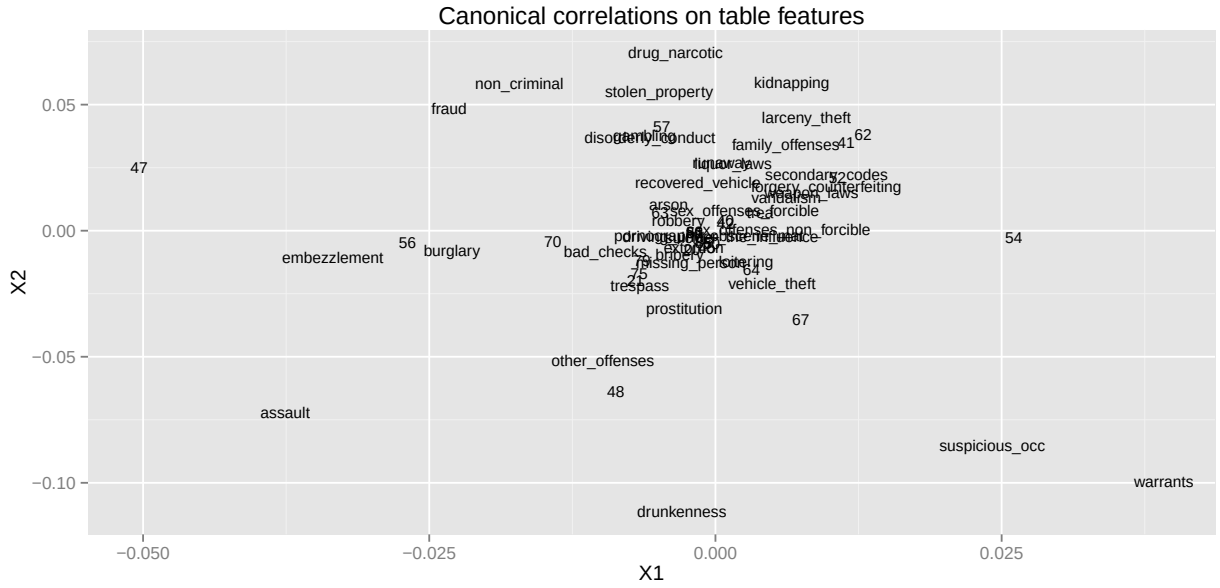



license_57
license_79
license_41
license_86
license_42
license_75
license_21
license_20
license_48
license_47
license_70
license_51
driving_under_the_influence
disorderly_conduct
drunkenness
liquor_laws
loitering
drug_narcotic
license_67
family_offenses
license_40
license_63
license_85
license_45
license_64
license_50
license_52
license_54
license_62
license_56
asian
Unemploy_p
black
hispanic
Pop2010
white
college
med_income
med_income
college
white
Pop2010
hispanic
black
Unemploy_p
asian
license_56
license_54
license_62
license_52
license_50
license_64
license_45
license_85
license_63
license_40
family_offenses
license_67
drug_narcotic
loitering
liquor_laws
drunkenness
disorderly_conduct
driving_under_the_influence
license_51
license_70
license_47
license_48
license_20
license_21
license_75
license_86
license_42
license_41
license_79
license_57



Canonical Correlation Analysis —

- Dimension reduction method popular when we have multiple tables. It's a generalization of PCA applicable when we have several tables
- Interpretation
- Satisfying that crimes that seem similar to each other are labeled that way
- Can look at tracts that are overrepresented in some kinds of crimes than others
- Doesn't seem to really associate with either income or unemployment. There is some unknown variation in tracts that drives this projection, but we haven't found the feature (or groups of features) driving that, at least not at this more cursory analysis...

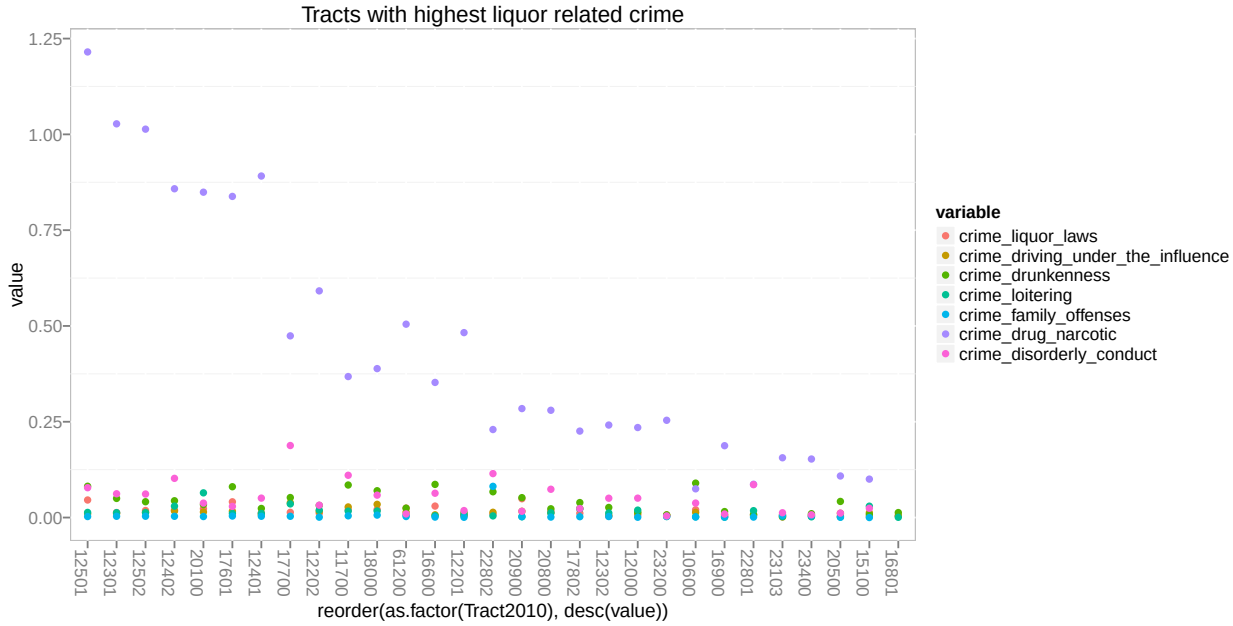


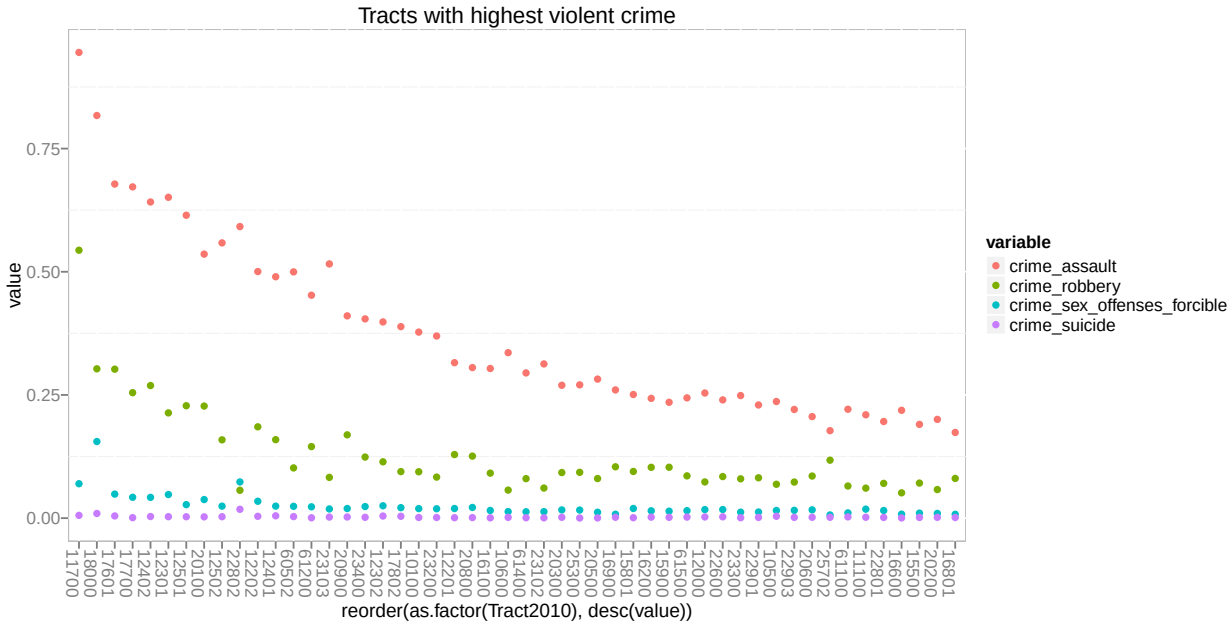
Detection of Anamolous Tracts —

Highly ranked —

Crime —

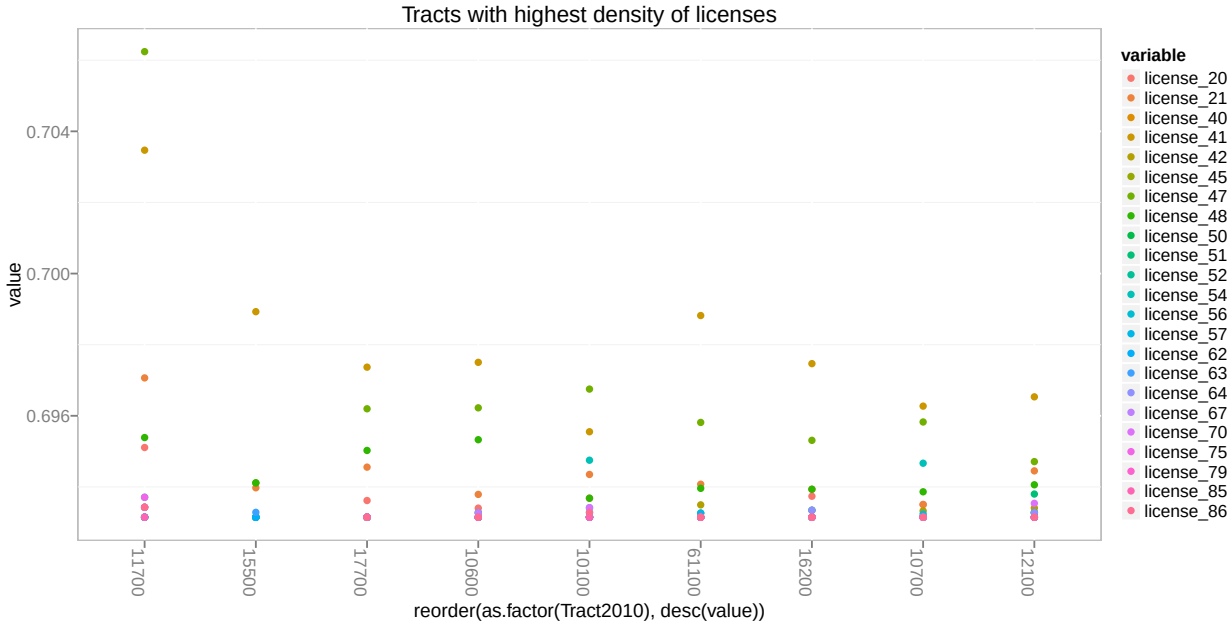
- What are the tracts that have the highest violent crime, liquor related crime





Liquor —

- What are tracts with the highest liquor store densities, overall
- Just on site
- Just off site



Matching: Comparing different distances —

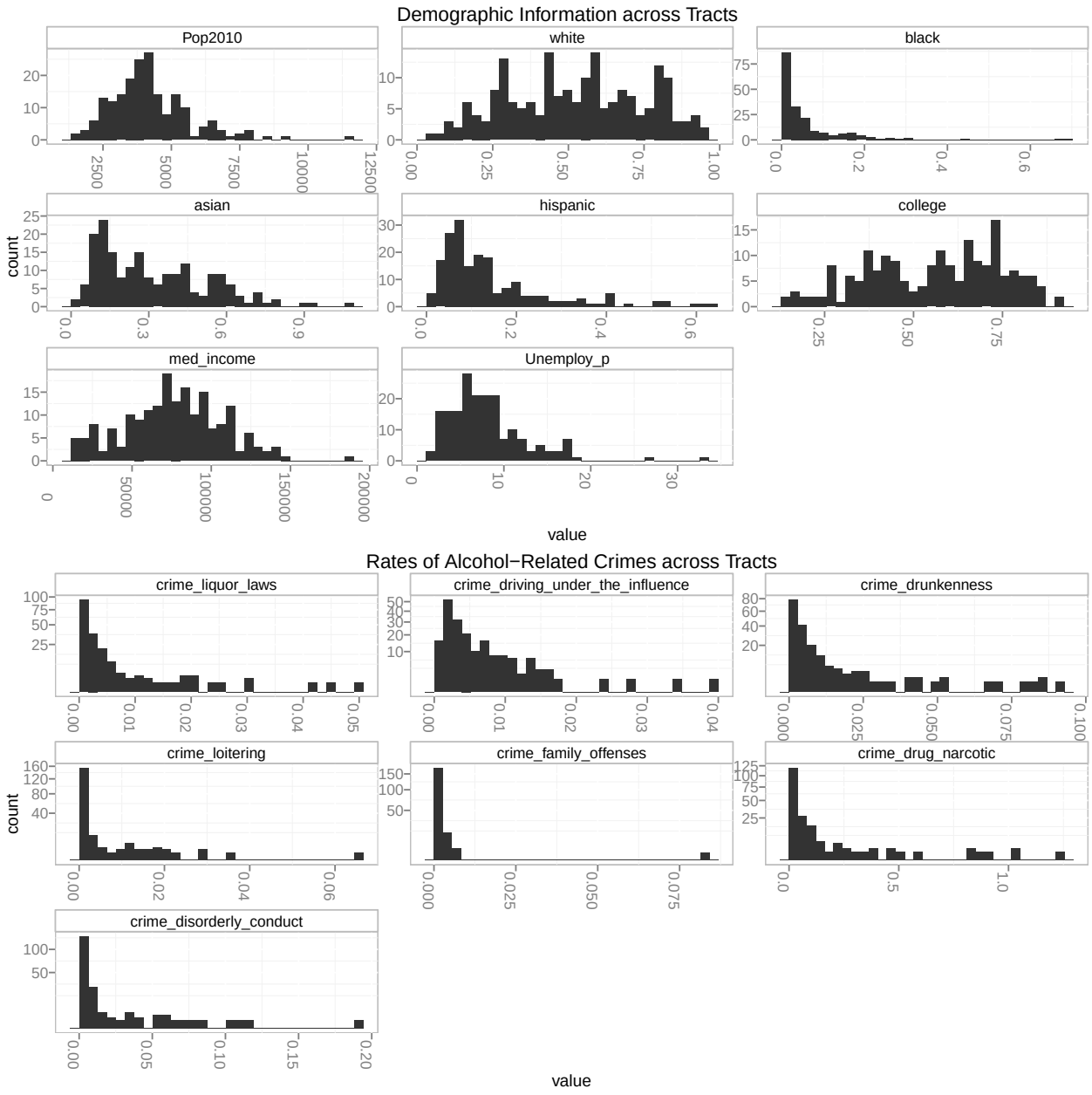
- When do two tracts seem very similar according to one data set but very different according to another?
- What are these tracts?

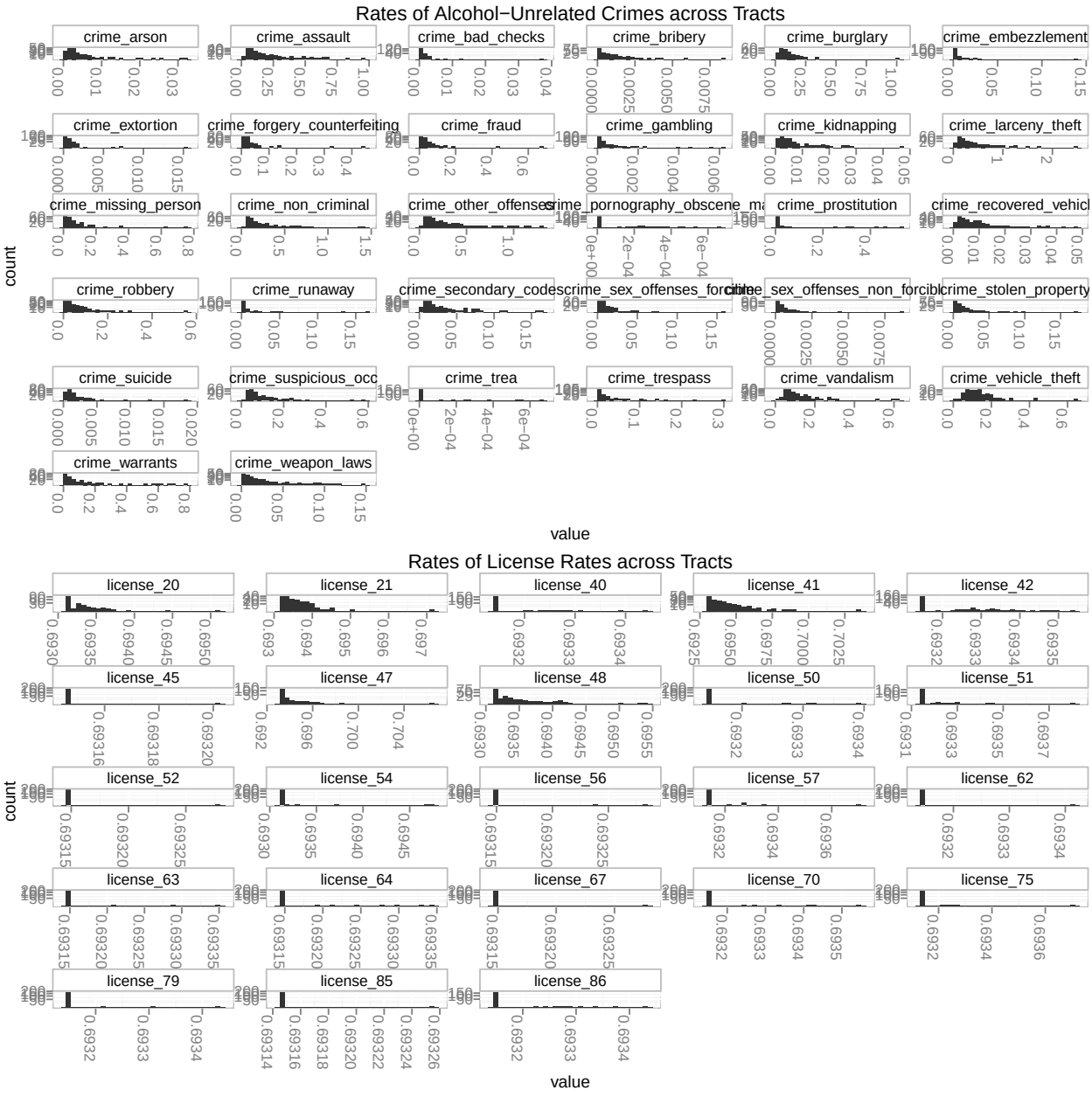
Limitations —

- Sales level information
- Proxies for patrolling bias
- Integration of 311 data
- Incorporation of spatial distance function

Appendix

Histograms of measured Variables across tracts




Rates of Alcohol−Unrelated Crimes across Tracts
crime_arson
crime_assault
crime_bad_checks
crime_bribery
crime_burglary
crime_embezzlement
crime_extortion
crime_forgery_counterfeiting
crime_fraud
crime_gambling
crime_kidnapping
crime_larceny_theft
crime_missing_person
crime_non_criminal
crime_other_offenses
crime_pornography_obscene_ma
crime_prostitution
crime_recovered_vehicl
crime_robbery
crime_runaway
crime_secondary_codes
crime_sex_offenses_forcible
crime_sex_offenses_non_forcibl
crime_stolen_property
crime_suicide
crime_suspicious_occ
crime_trea
crime_trespass
crime_vandalism
crime_vehicle_theft
crime_warrants
crime_weapon_laws
count
value

Rates of License Rates across Tracts
license_20
license_21
license_40
license_41
license_42
license_45
license_47
license_48
license_50
license_51
license_52
license_54
license_56
license_57
license_62
license_63
license_64
license_67
license_70
license_75
license_79
license_85
license_86
count
value