MACS 40800: Unsupervised Machine Learning - Problem Set #1

Erika Tyagi

10/9/2019

Exploration & Computation

Question 1

I'm using block group-level data for Cook County, IL on evictions and eviction filings, made available from The Eviction Lab. These data are provided annually between 2006 and 2016. Note that each row in the dataset corresponds to a single block group / year combination. The data are available for download here: https://evictionlab.org/get-the-data/.

```r
# limit to Cook County from 2006 - 2016
chi_evictions <- read.csv("block-groups.csv") %>%
    filter(parent.location == "Cook County, Illinois",
           year %in% (2006:2016)) %>%
    select(GEOID, year, evictions, eviction.filings)
```

Question 2

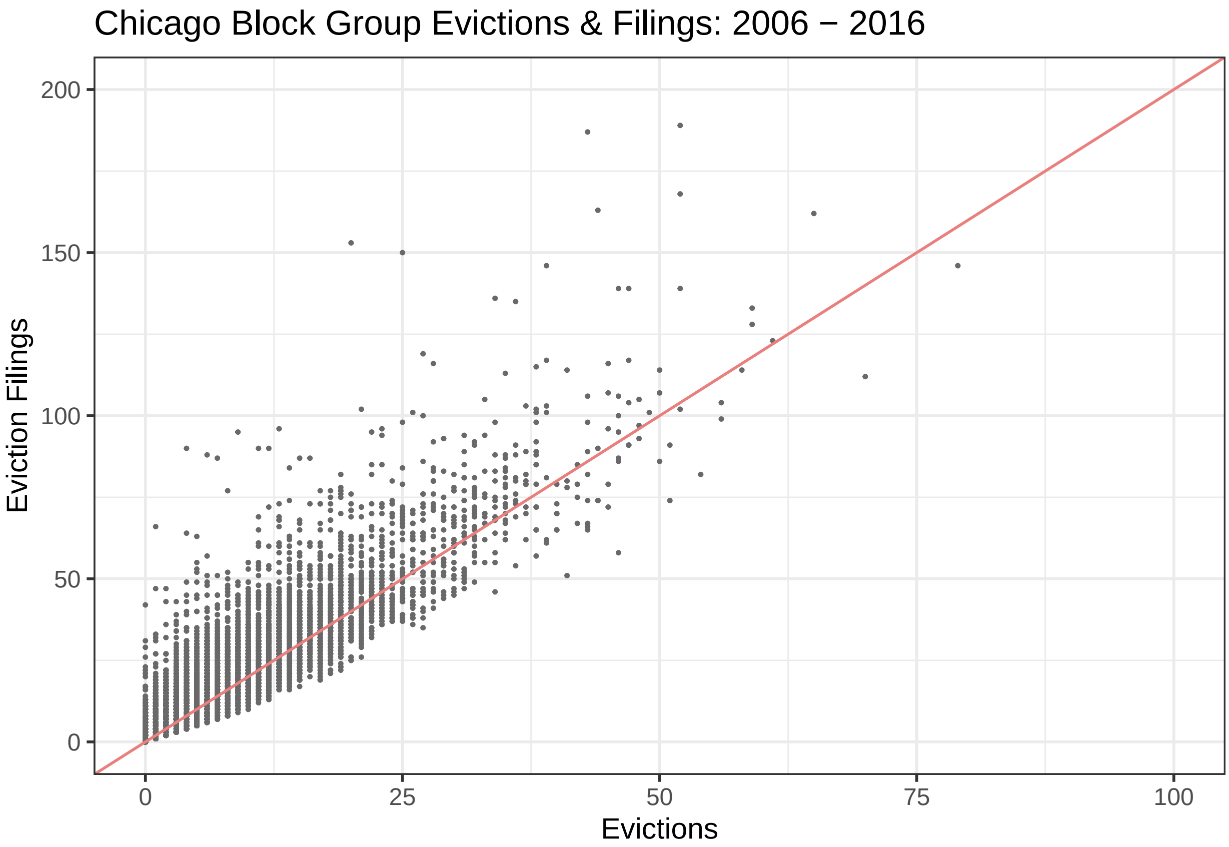
I'm creating a scatterplot to visualize the relationship between the number of eviction filings and the number of actual evictions.

An eviction filing is defined as "the result of a landlord filing a case in court to have a tenant removed from a property," whereas an eviction "happens when a landlord expels people from property he or she owns." One of the major goals of policy advocates is to increase the delta between filings and actual evictions – or provide tenants with the resources (typically legal representation) to successfully fight back against a wrongful eviction filing. Thus, getting a sense of the relationship between these two variables across block groups is meaningful.

Question 3

Unsurprisingly, there is a positive correlation between eviction filings and evictions (i.e., block groups with a high number of eviction filings also had a high number of evictions). The vertical line represents a 2:1 relationship where half of all filings led to actual evictions. Points above the line reflect blocks where fewer than half of all evictions materialized into evictions; points below reflect blocks where more than half of all evictions did so.

```r
# scatter plot of evictions and filings
chi_evictions %>%
    ggplot(aes(x = evictions, y = eviction.filings)) +
    geom_point(size = 0.3, color = "#696969") +
    xlim(0, 100) + ylim(0, 200) +
    geom_abline(intercept = 0, slope = 2.0, color = "#e9807d") +
    labs(x = "Evictions",
         y = "Eviction Filings",
         title = "Chicago Block Group Evictions & Filings: 2006 - 2016") +
    theme_bw()
```



Question 4

```
round(
    stat.desc(
        chi_evictions %>%
            select(evictions, eviction.filings)),
      1)
```

```
##               evictions eviction.filings
## nbr.val         43923.0          43923.0
## nbr.null        12042.0           4629.0
## nbr.na              0.0              0.0
## min                 0.0              0.0
## max                79.0            189.0
## range              79.0            189.0
## sum            150903.0         383691.0
## median              2.0              5.0
## mean                3.4              8.7
## SE.mean             0.0              0.1
## CI.mean.0.95        0.0              0.1
## var                24.0            130.4
## std.dev             4.9             11.4
## coef.var            1.4              1.3
```

Question 5

The mean number of eviction filings (per block group per year) is almost 9, while the mean number of evictions is just over 3. As both variables have a strong right skew, the medians for both are lower than the means – 5 for filings and 2 for evictions. Most blocks had very few evictions (e.g., the bottom quartile all had 0 evictions).

This is particularly relevant when considering the importance of data quality. Per The Eviction Lab's FAQs (https://evictionlab.org/help-faq/#data-source), the data are collected from a combination of sources (local court records and proprietary eviction records datasets) – and ensuring these data are perfectly complete and standardized is an impossible task. As a result, using these data to authoritatively say that a block with 2 evictions is substantively different from a block with 0 evictions may not be valid – even though this would reflect jumping from the 0th to the 50th percentile.

Critical Thinking

Question 1

Both visual and numeric EDA can be useful in identifying outliers, trends, and patterns in data. Numeric analysis can be particularly useful for providing precise and concrete 'apples to apples' comparisons (e.g., comparing the mean / median / variance across groups, etc.). Visual analysis is well-suited for revealing patterns and general trends that can't be reduced to the 'science' of a precise number, but reflect the 'art' of human perception. Conveying distributions, trends, relative scales, and other more subjective differences is often easier for humans to digest visually – although, mis-representing these concepts visually is also an easy pitfall.

Question 2

Example 1: https://badvisualisations.tumblr.com/image/183585563181
Put simply, this visualization has too much going on to be particularly useful. It's unclear where to start digesting its content, what the meaningful takeaways are, how to make sense of the layers of legends (I think there are at least 3?), etc. I also don't think any amount of zooming in will make the text legible while also allowing for the relevant relationships to still be understood.

Example 2: https://badvisualisations.tumblr.com/image/183573397446
This visualization make the already undesirable pie chart even worse – most glaringly, the total summing to 33.52% isn't intuitive. It's also unclear what exactly the chunks of the pie represent, and how these relate to the quotes to the left.

Question 3

Example 1: https://www.nytimes.com/interactive/2019/06/18/upshot/cities-across-america-question-single-family-zoning.html
I love this collection of maps because they're simple, intuitive, visually striking, and effectively (and efficiently) get across a simple point – a lot of land in major cities is zoned for single-family housing. I think the fact that it takes just a second for a reader to come away with this point makes them really nice complements to the article.

Example 2: http://apps.urban.org/features/mapping-americas-futures/
The map itself is simple, uses visually engaging colors, and conveys meaningful information without being overly complex. I particularly like the 'click to discover more' option, which allows a viewer to simultaneously (1) quickly get a 'big picture' national view, but (2) allow an interested individual to dive into dive into the weeds of a particular area. I think that separating these two components (but making transitioning between them easy) is a particularly effective technique.

Question 4

EDA can be useful across nearly all steps in the research process – but is particularly useful during the initial exploratory phase of research. Specifically, it's useful for understanding and validating assumptions and trends in data, identifying patterns and hypotheses, and formulating research questions. When working with real data, EDA is a necessary step for diagnosing data quality. When working on computational (or coding-intensive) projects, integrating EDA techniques into a workflow can also be really useful as a 'sanity check' when performing complex data wrangling or programming tasks.

Question 5

Confirmatory data analysis can be broadly thought of as formal hypothesis testing, while exploratory analysis reflects using data to suggest hypotheses to test. An example of confirmatory analysis can be running a regression with an already formulated hypothesis used to assess the statistical significance of [x] on [y] holding [z] constant. Exploratory analysis would be a less structured way of finding ways to examine various relationships, with the goal of understanding what a meaningful [x], [y], and [z] would be.