Optimizing Proactive Inspections

McCourt School of Public Policy
Data Science for Public Policy Program

1789

GEORGETOWN UNIVERSITY

Acknowledgements

DCRA

- Director Ernest Chapprah
- Tomash Bukowiecki
- Anthony Clark
- Abiola Ijaola
- Peter Nyabagosi
- Garret Whitescarver

Georgetown University

- NaLette Brodnax
- Christopher Murphy
- Jinia Sarkar
- Heidi Tseu
- Massive Data Institute

McCourt DSPP Students

- Jean Landry Binam Keyanfe
- Lixue Chen
- Chandler Dawson
- Zachary Gozlan
- Douglas Hummel-Price
- Yining Li
- Ziqi Niu
- Madeline Pickens
- Andrea Cristina Quevedo Acuna
- George Schoeffel
- Harsh Sharda
- Neha Tiwari
- Lily Yue

Project Timeline

August 28 — Meeting with DCRA Office of Data Innovation

September 23 — Meeting with Garret Whitescarver, Deputy Chief Building Official

October -- November — Three in-class labs, exploring the data and deriving insights

November 18 — Insight Competition

Overview

How can DCRA optimize proactive inspections of licensed properties?

How Properties are Currently Selected for Inspection

- Licensed properties with more than three units
- Properties not inspected in the last four years
- Random selection by ward

Potential Issues with Current Selection Method

- DCRA resources may not be focused on problem areas
- Owners have minimum incentive to improve properties if they have been inspected in the last four years

Inspection Violation Rate as Measure of DCRA Success

- Inspection violation rate is defined as:

$$\frac{\mathit{Initial\ Inspections\ with\ Violations}}{\mathit{Total\ Initial\ Violations}}$$

- In the following example, the inspection violation rate is 6 / 12 = 50%

Compliance Rate as Measure of DCRA Success

- NOV Compliance Rate is defined as:

$$\frac{Total\ Violations\ Corrected\ at\ Follow{-}Up}{Total\ Follow{-}Up\ Inspections}$$

- In the following example, the compliance rate is 10 / 12 = 83 %

Data Overview

- DCRA data
  - Proactive inspections
  - Complaint-based inspections
  - Violation severity data (for both proactive and complaint-based)
  - Residential license data

- Computer-Assisted Mass Appraisal data
  - Property characteristics for commercial, condo, residential

Proposed Selection Approaches

Choose properties on basis of:

- Violation history
- Complaint severity
- Property characteristics
- License characteristics

Insights (1/3)

Analysis of Complaint Inspection Data: Inspection History

Initial Violation Rate (IVR) Trends

Year	Proactive Inspections	Inspection Violation Rate
2016	561	53.8%
2017	2,736	70.6%
2018	2,267	56.0%
2019	2,231	52.7%

Highest IVR in 2017

Decreasing IVR after 2017

Why?

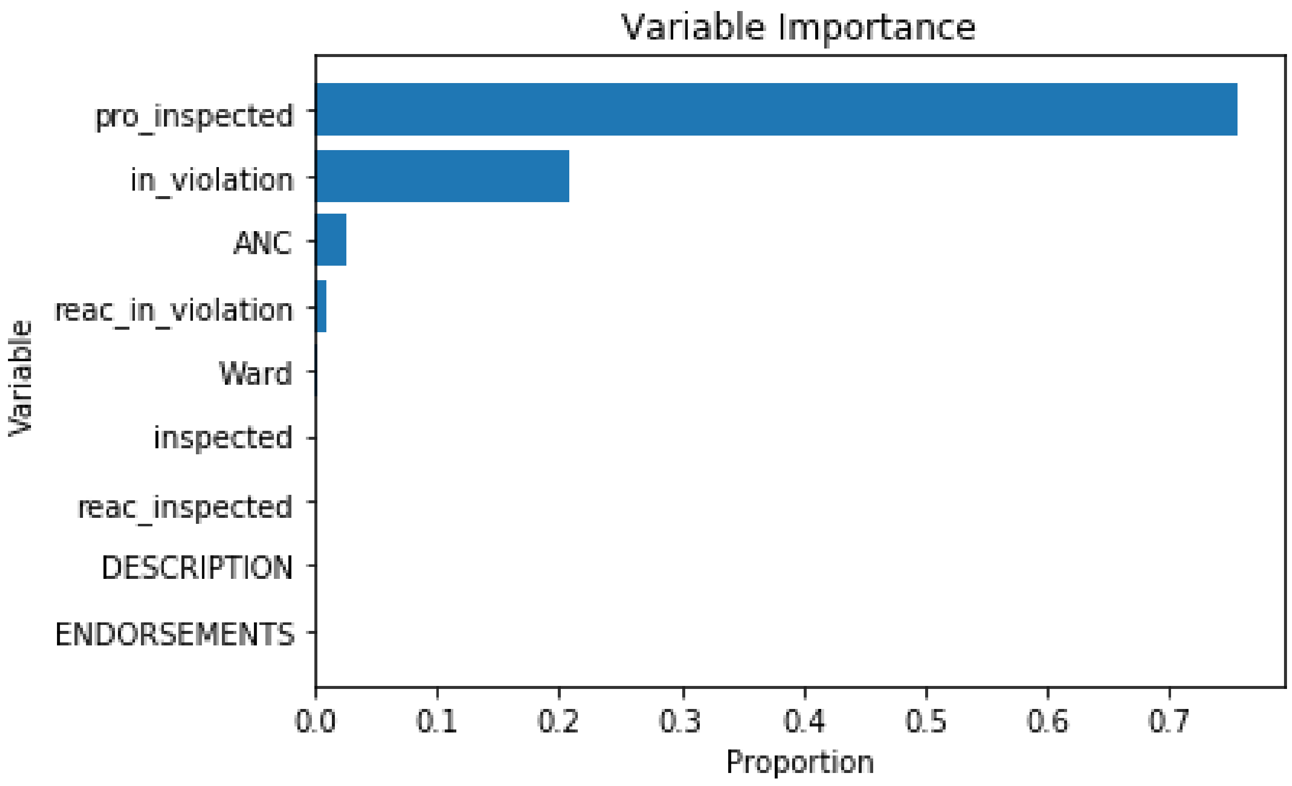
Higher IVR when properties with complaints inspected

Year	Proactive Inspections	Inspection Violation Rate	Complaint Inspections
2016	561	53.8%	1,815
2017	2,736	70.6%	7,491
2018	2,267	56.0%	6,637
2019	2,231	52.7%	6,637

Year 2017 had highest IVR and **highest number of address IDs in common** between proactive and reactive inspections

Insight: Properties with history of complaints more likely to have violations

Use complaint history to predict violations



Key takeaway: Use inspection history to assign proactive inspections

- Variables
  - Reactive inspection of property due to a complaint (since 2015)
  - Proactive inspection of property (since 2015)
  - Property was inspected whether proactive or reactive (since 2015)
- Result: 98% prediction accuracy

Insights (2/3)

Analysis of Complaint Inspection Data:

Complaint Severity

Text Analysis Goals

1. Comparing -- Examine similarity

2. Clustering -- Identify groups

3. Summarizing -- Show important topics or themes

We do this by turning our words into tokens, or data objects that hold information for each word

Text Analysis Example

Sample Text:

The DCRA data had about eight thousand observations.

Step 1: Clean

the dcra data had about eight thousand observations

Step 2: Remove Stopwords

dcra data eight thousand observations

Step 3: Tokenize

[‘dcra’, ’data’, ’eight’, ’thousand’, ’observations’]

DCRA Example

Example Complaint:

Rust on floor and railing on the 4th floor with water damage in the kitchen with peeling paint on the wall

Management: Random Land Company

123-456-7890 Mr. John Doe

Anonymous Caller

Reduce complexity of complaint text

Rust, floor, railing, 4th, floor, water, damage, kitchen, peeling, paint, wall, Management, Random Land Company, 123-456-7890, Mr. John Doe, Anonymous, Caller

Steps

1. Remove Numbers
2. Remove Names
3. Remove unnecessary information

Outcome: rust, floor, railing, water, damage, kitchen, peeling, paint, wall

Next Step: Analyze terms to identify topics that appear frequently

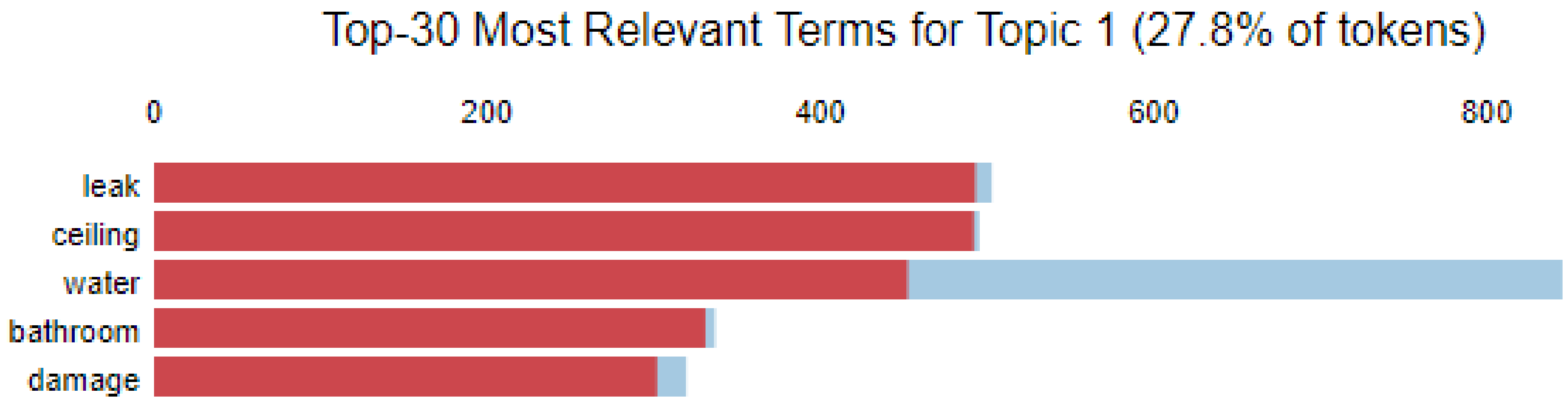
Interpreting the Topics

	Word 1	Word 2	Word 3	Word 4	Word 5
Topic 1	basement	flood	sewer	rain	electrical
Topic 2	need	bug	home	vacant	bed
Topic 3	door	defective	work	paint	window
Topic 4	infestation	multiple	water	work	mouse
Topic 5	leak	ceiling	water	bathroom	damage
Topic 6	heat	dcra	jep	possible	inspect
Topic 7	water	hot	remove	electricity	emergency

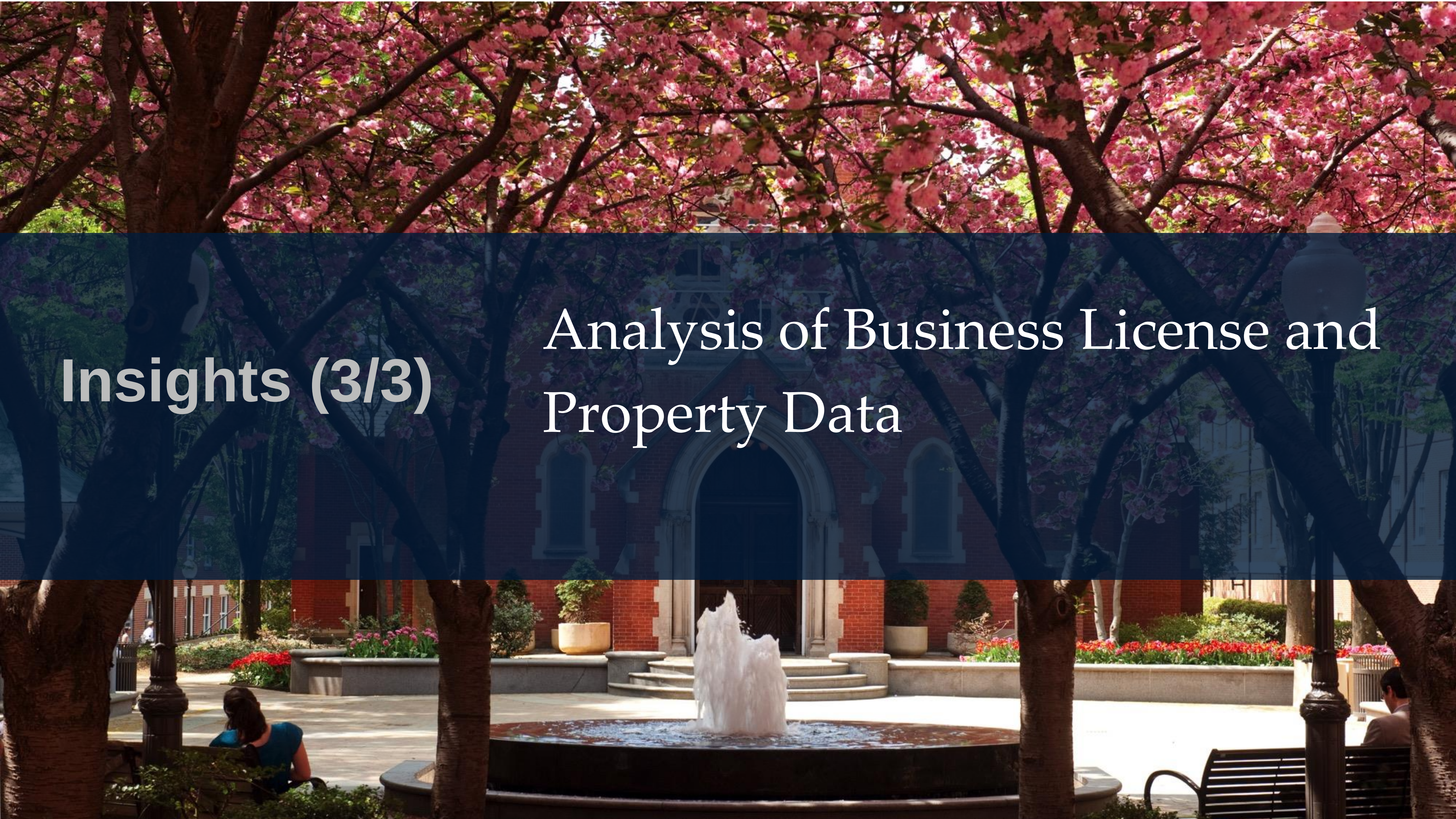
Categorizing Complaints

Topics	% of Complaints
Water leaks coming from a ceiling or outside causing structural issues like collapsed ceilings	28%
Degradation of the condition of the living space	12%
Broken safety equipment or entry point, like windows and doors	16%
Infestation of bugs or rats within a unit	17%
Water issues with the roof or basement of a building, including flooding	6%
Issues regarding the heat of a living space	12%
Problems comings from utilities	9%

Visualizing the Topics



Key takeaway: Utilize text in complaints in combination with inspection history to identify properties with a history of certain types of complaints



Insights (3/3)

Analysis of Business License and Property Data

Techniques & Datasets

The business license/property analyses use the following techniques:

- Principal Components Analysis (PCA),
- Regression (e.g., OLS and Logistic Regression)
- Predictive Modeling (e.g., Random Forest)

The analyses use the following datasets:

- Proactive Inspections
- Residential Licenses
- Violation Data (Count, Fines, Severity)
- Computer-Assisted Mass Appraisal

Result 1: Actual Year Built

The selection algorithm could weight properties to more frequently target buildings based on construction and renovation dates

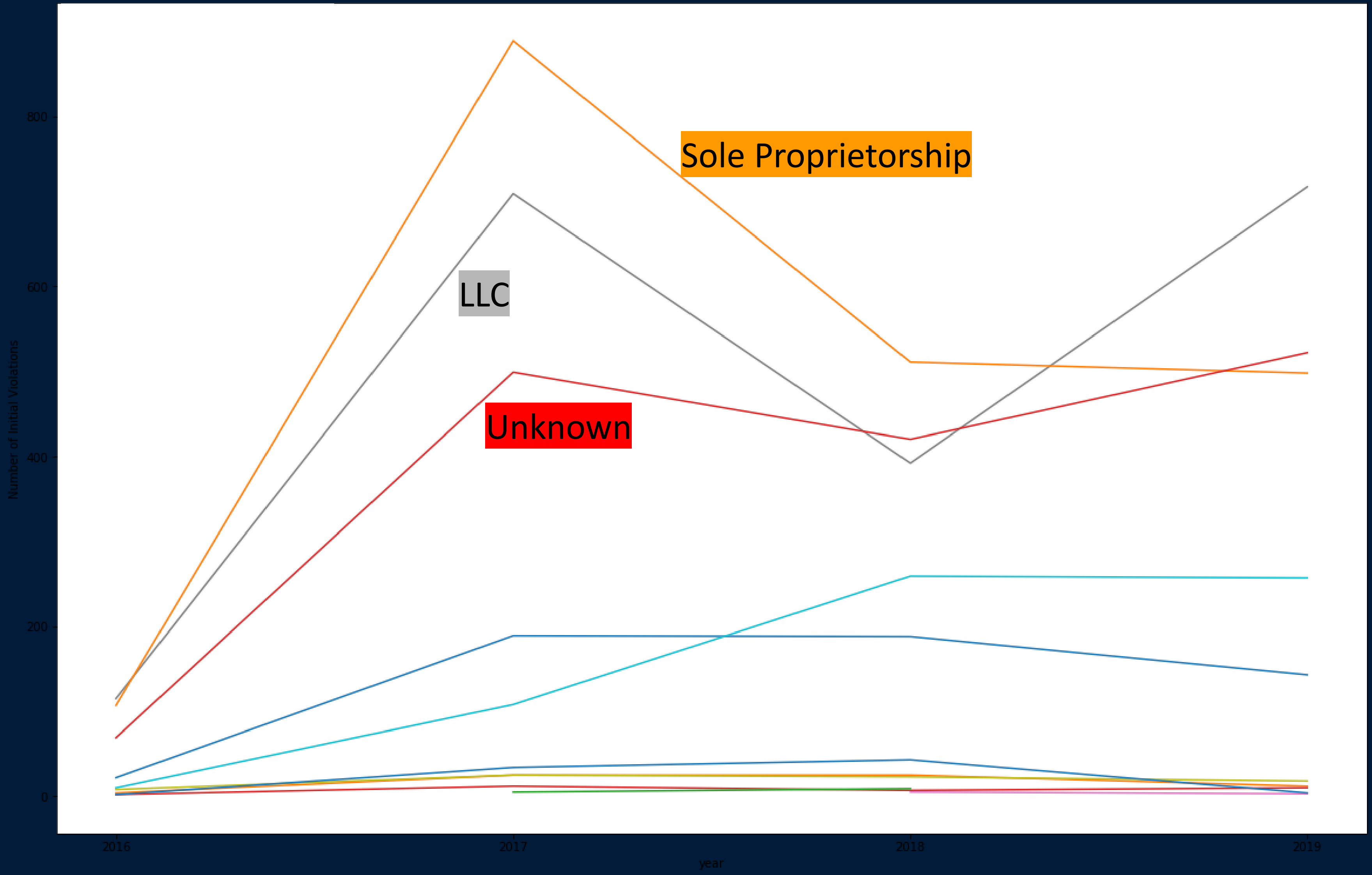
- AYB-- actual year built
- EYB-- effective year built (factors in renovation)

Variable	Importance
AYB	0.169473
EYB	0.137872
ROOMS	0.106589
SALE_NUM	0.094474
BEDRM	0.084557
BATHRM	0.070715
STORIES	0.069339
AC	0.057188
FIREPLACES	0.056283
KITCHENS	0.052981
HF_BATHRM	0.051654
NUM_UNITS	0.048875

Result 2: Ownership Type

Top 3 categories for the number of violations found:

- Sole Proprietorship
- Limited Liability Partnership
- Unknown



Result 3: Business License Description

Compared to apartment, cooperative association is more likely to have 30-day violations, but is less likely to have 1-day violations.

Two family rentals are less likely to have 1-day violations compared to apartment.

OLS Regression Results for 30-day violations

	coef	std err	t	P>|t|	[0.025	0.975]
Cooperative Association	2.5907	0.759	3.413	0.001	1.102	4.079
One Family Rental	0.5995	0.597	1.005	0.315	-0.570	1.769
Two Family Rental	0.0636	0.619	0.103	0.918	-1.150	1.277

OLS Regression Results for 1-day violations

	coef	std err	t	P>|t|	[0.025	0.975]
Cooperative Association	-1.2405	0.429	-2.894	0.004	-2.081	-0.400
One Family Rental	0.0106	0.337	0.031	0.975	-0.650	0.671
Two Family Rental	-1.0709	0.350	-3.064	0.002	-1.756	-0.386

Result 3: Business license status matters

Compared to active status, properties with expired licenses are more likely to have 30-day violations.

Compared to active status, properties with cancelled licenses are less likely to have 1-day violations.

OLS Regression Results for 30-day violations

	coef	std err	t	P>|t|	[0.025	0.975]
Cancelled	-0.1904	0.125	-1.518	0.129	-0.436	0.056
Expired	0.3424	0.118	2.892	0.004	0.110	0.575
Others	1.1092	0.426	2.602	0.009	0.273	1.945

OLS Regression Results for 1-day violations

	coef	std err	t	P>|t|	[0.025	0.975]
Cancelled	-0.2033	0.071	-2.868	0.004	-0.342	-0.064
Expired	0.1274	0.067	1.906	0.057	-0.004	0.259
Others	1.0425	0.241	4.331	0.000	0.570	1.515

Next Steps

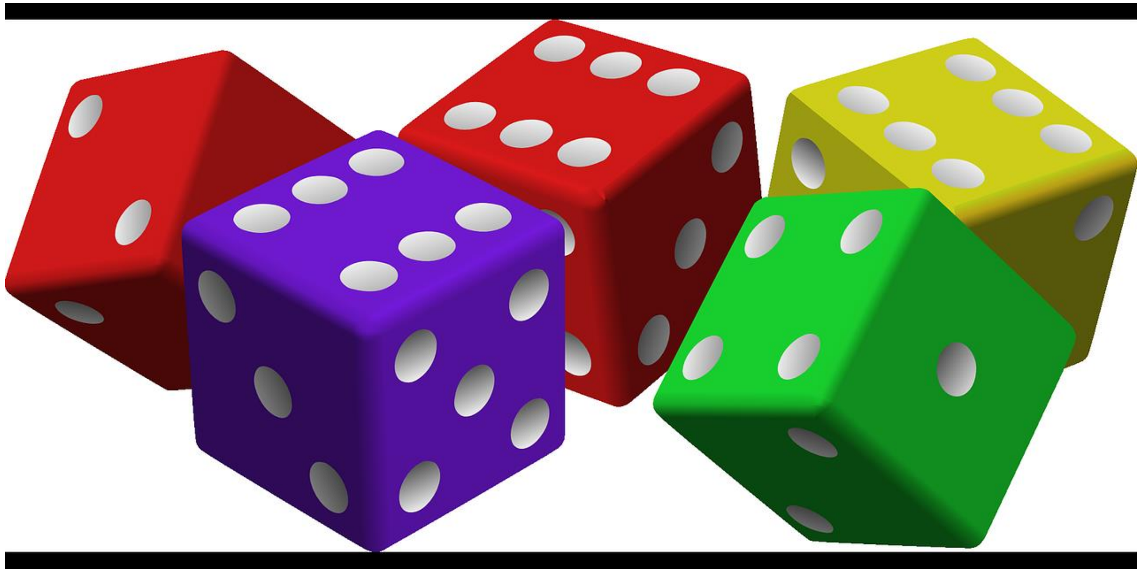
Implementing a New Proactive Inspection Selection Algorithm

Probability Fundamentals

The **probability** that event X will occur is a numerical measure that represents how likely or unlikely a given event is to occur.

For example:

With a fair six-sided die, each side is equally likely to come up, so the probability for each is 1 in 6.



Unweighted Probability

Think of putting names into a hat, and then pulling one name out. Each name is equally likely to be drawn, so we consider this probability to be **unweighted.**

Unweighted probabilities work well in situations where each “name in the hat” should have an equal chance of being selected.

Weighted Probabilities

Weighted probabilities, by contrast, account for situations in which some names should be more likely than others.

For example, the NBA draft uses a weighted lottery system to give teams that performed poorly the previous season a greater chance to draft better players. This prevents a dominate team from building on their advantage.

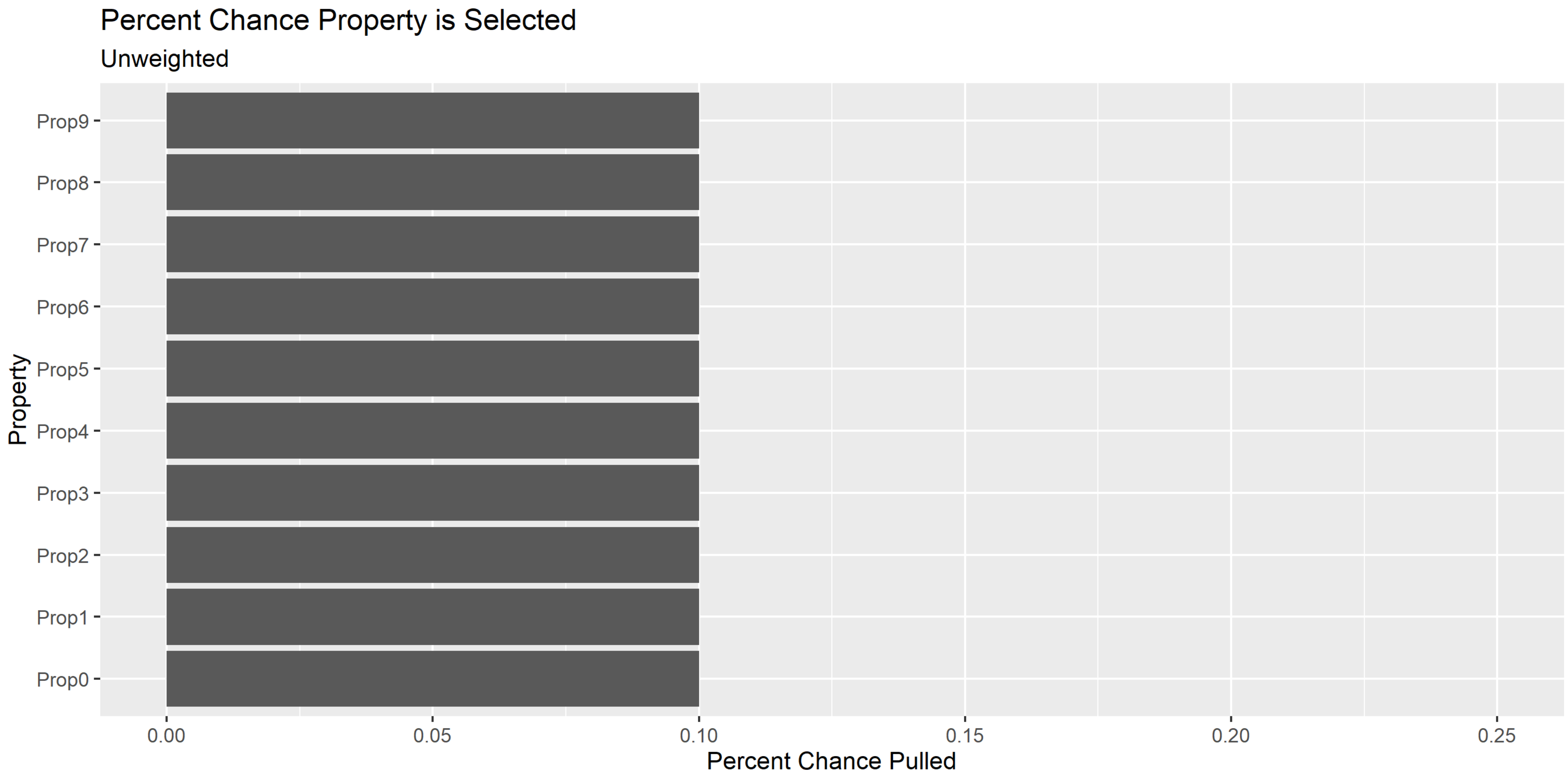
For the names in a hat situation, this is analogous to adding some names to the hat more frequently than others.

Weighting in DCRA Context

Let's say DCRA is selecting one property to inspect from a group of ten properties.

When the selection is random, each one has an equal chance of being selected.

Property	Chance Property is Selected
Property 0	10%
Property 1	10%
Property 2	10%
Property 3	10%
Property 4	10%
Property 5	10%
Property 6	10%
Property 7	10%
Property 8	10%
Property 9	10%


Percent Chance Property is Selected
Unweighted
Property
Prop9
Prop8
Prop7
Prop6
Prop5
Prop4
Prop3
Prop2
Prop1
Prop0
0.00
0.05
0.10
0.15
0.20
0.25
Percent Chance Pulled

Weighting in DCRA Context

But if we factor in what we know based on past outcomes, we can make it so that properties that **look more like** properties where violations were found are more likely to be selected for the inspection.

Let's say here, we know that certain properties are in zip codes where DCRA finds violations frequently; older buildings are more likely to have violations; and certain homes are built from cheaper building materials.

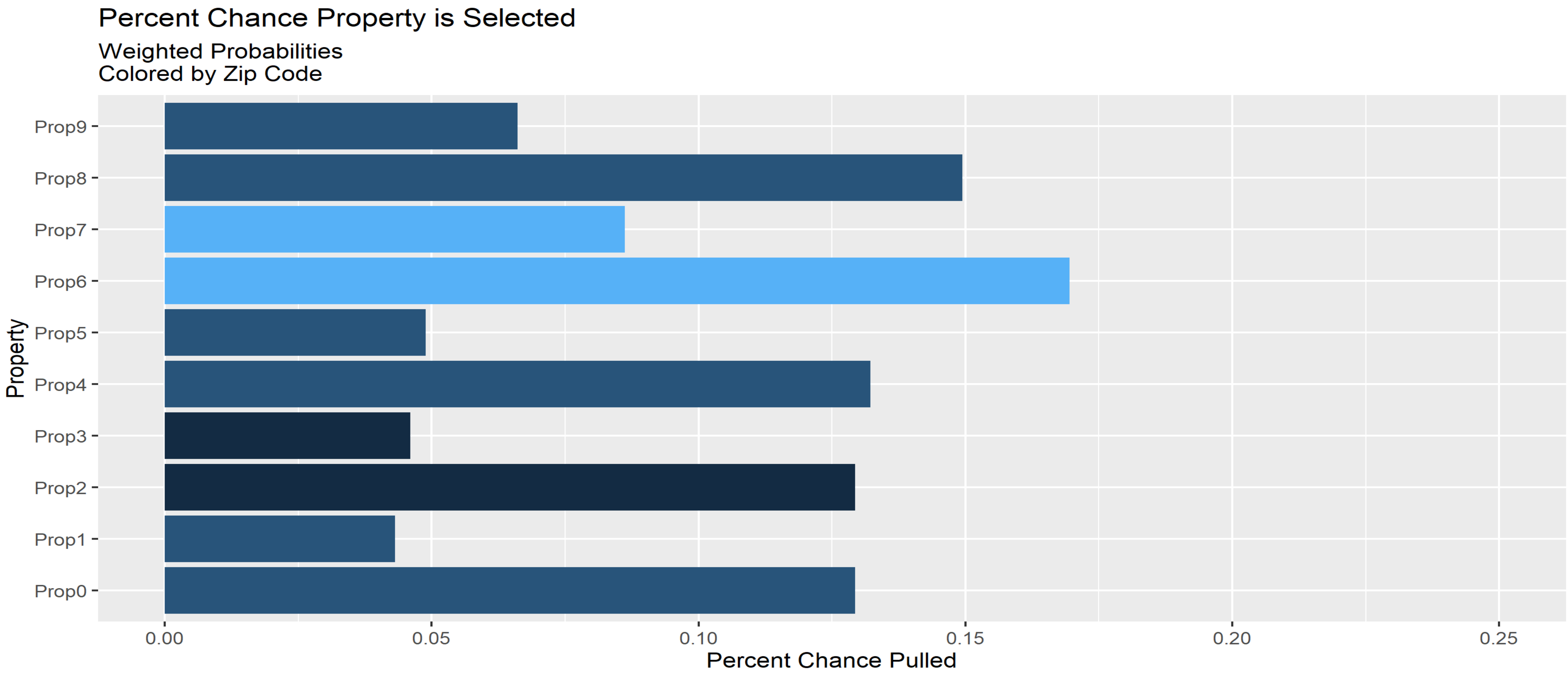
Property	Zip Code History	Year Built	Building Material
Property 0	Normal	1900	Standard
Property 1	Normal	1900	Cheap, Flammable Wood
Property 2	Few Violations	1950	Standard
Property 3	Few Violations	1960	Cheap, Flammable Wood
Property 4	Normal	1910	Standard
Property 5	Normal	1925	Cheap, Flammable Wood
Property 6	Many Violations	1945	Standard
Property 7	Many Violations	1950	Cheap, Flammable Wood
Property 8	Normal	1970	Standard
Property 9	Normal	1980	Cheap, Flammable Wood

Weighting in DCRA Context

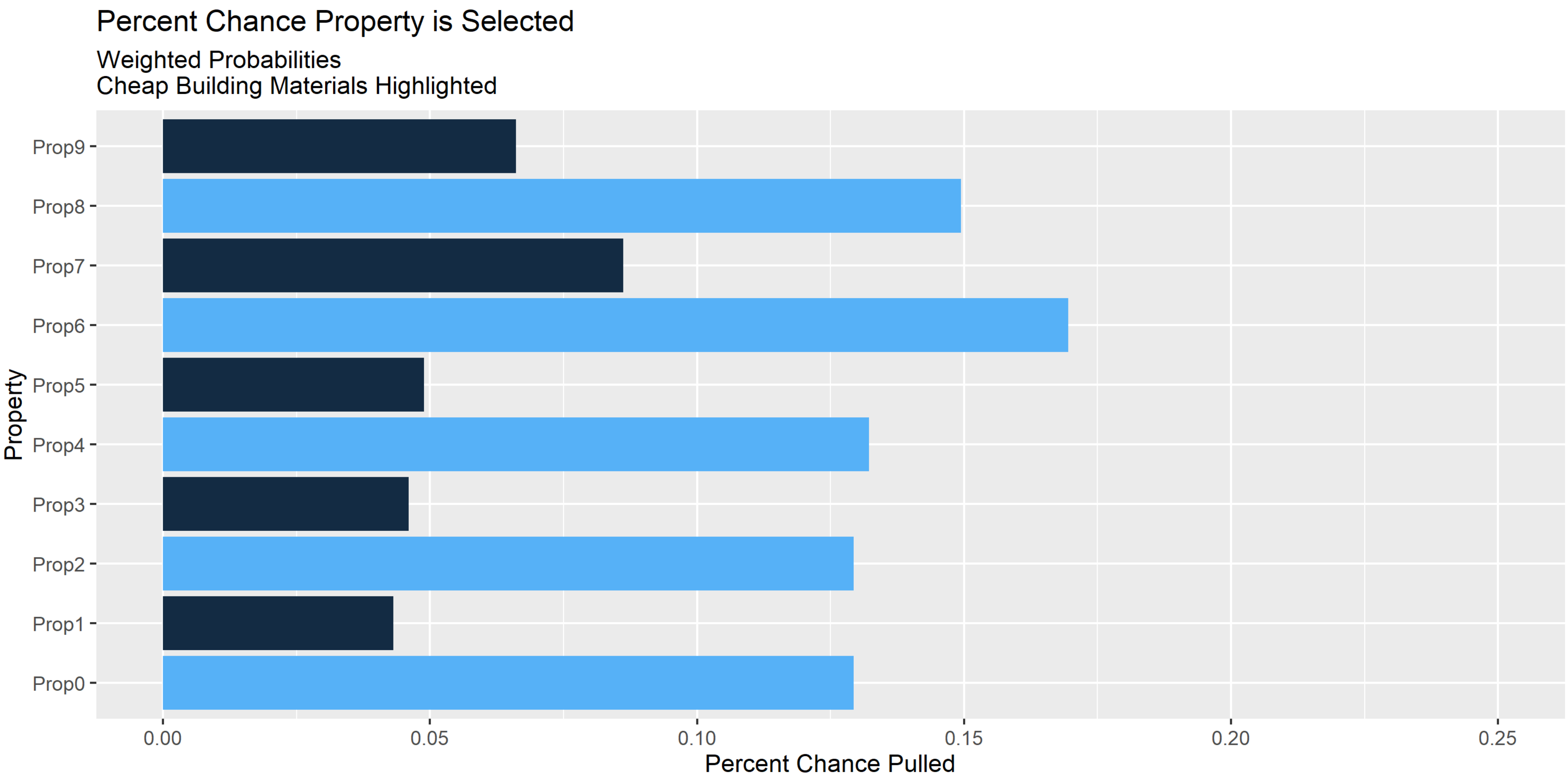
By weighting these characteristics, we can make it so that properties with more "warning signs" are more likely – but still not guaranteed - to be chosen for inspection:

Property	**Weighted** Chance Property is Selected	Zip Code History	Year Built	Building Material
Property 0	**12.9%**	Normal	1900	Standard
Property 1	**4.3%**	Normal	1900	Cheap, Flammable Wood
Property 2	**12.9%**	Few Violations	1950	Standard
Property 3	**4.5%**	Few Violations	1960	Cheap, Flammable Wood
Property 4	**13.2%**	Normal	1910	Standard
Property 5	**4.9%**	Normal	1925	Cheap, Flammable Wood
Property 6	**17.0%**	Many Violations	1945	Standard
Property 7	**8.6%**	Many Violations	1950	Cheap, Flammable Wood
Property 8	**14.9%**	Normal	1970	Standard
Property 9	**6.6%**	Normal	1980	Cheap, Flammable Wood

Using slightly more advanced techniques, we can make judgments between the weights, such that, say, zip code sways things some…



…but an unsafe building material sways things **more.**



Summary of Recommendations

Apply weighted probabilities to properties on the basis of

1. History of prior violations (proactive and reactive)
2. History of complaints with particular topics or keywords
3. Older construction or recent renovation
4. Expired license status
5. Owned by a sole proprietor or LLC
6. Licensed as an apartment or cooperative association

Questions?