# Exploring Housing, Homelessness and Public Syringe Recovery

JOHN FERRARA

FINAL PROJECT (DATA 606 & DATA 607)

2024-12-08

# Research Questions

**Does housing impact this seemingly separate public health and safety issue of public/ improper syringe disposal?**

**Question 1:** Does the Number of Housing Related Construction Projects for the Previous 5 Years correlate with the Size of Adult Homeless Population?

- **Null Hypothesis:** The number of housing related construction projects for the preceding 5 year window period does not correlate with the adult homeless population.
- **Alternative Hypothesis:** The number of housing related construction projects for the preceding 5 year window period does correlated with the adult homeless population.
- **Variables:**
  - *Independent*: The independent variable in this question is the completed housing-related construction projects completed in the preceding 5 year window from when the homeless population was counted.
  - *Dependent*: The dependent variable in this question is the Adult homeless population.

**Question 2:** Does the size of the Adult Homeless Population correlate with the Number of Used Syringes recovered by NYC Parks?

- **Null Hypothesis:** The size of the Adult homeless population does not correlate with the number of used syringes recovered in NYC Parks.
- **Alternative Hypothesis:** The size of the Adult homeless population does correlate with the number of used syringes recovered in NYC Parks.
- **Variables:**
  - *Independent*: The independent variable in this question is the adult unhoused population.
  - *Dependent*: The dependent variable in this question is the total syringes collected from NYC Parks and public safe disposal sites.

# Data Sources

## NYC Parks Syringe Collection Data

**Overview:**

- Collected by NYC Parks Department staff and non-profits.
- Tracks improperly discarded syringes in public parks.
- Data grouped by Parks Districts, each row represents daily totals.

**Key Assumptions:**

- Total ground syringes collected reflect improper disposal trends.
- Combining kiosk and ground syringe totals ensures comprehensive data.
- Syringe disposal habits may vary between housed and unhoused populations but are not explicitly categorized in the data.

**Source Format & Ingestion:**

- Data sourced as CSV files.
- Ingested using NYC Open Data API iteratively.

# Data Sources

## NYC Dept. City Planning Housing Database

**Overview:**

Data from NYC Department of City Planning (DCP).

Includes construction, renovation, and demolition job counts by geographic area.

Analysis focused on yearly counts of completed projects for Community Districts (2010–2020).

**Assumptions:**

Higher construction job counts correlate with improved housing conditions/availability.

Data does not differentiate between new buildings, alterations, or demolitions.

**Source Format & Ingestion:**

Data downloaded as a zip file.

Extracted and processed a single CSV file for Community District data.

# Data Sources

## NYC Parks Districts

**Overview:**

- Maps NYC Parks Districts to Community Districts.
- Parsing and flattening for multiple CD values per Parks District needed.
- Used to estimate syringe counts at the Community District level.

**Key Steps:**

- Syringe data joined with crosswalk to map syringe counts to Community Districts.
- A crosswalk table providing a unique value count for mapping.

**Source Format & Ingestion:**

- Directly ingested from a single CSV URL

# Data Sources

## NYC Dept. Homelessness Services Individual Census

**Overview:**

- Contains population counts for various types of shelters across NYC.
- Focused on adult-only shelter categories (e.g., Adult Family Shelter, Adult Shelter).

**Key Assumptions:**

- Intravenous drug users likely to be adults, not families with children.
- Excluded shelter counts for family-only categories.

**Source Format & Ingestion:**

- Data sourced from JSON API.
- Processed by summing relevant columns and averaging annual values for Community Districts.

# Ingestion Methods

**1. NYC Parks Syringe Collection Data**

- API ingestion with chunking (1,000-row increments).
- Combined multiple chunks into a single Data Frame.

**2. NYC Dept. City Planning Housing Database**

- Downloaded zip file from the NYC DCP website.
- Extracted and processed a specific CSV for Community District data.

**3. NYC Parks Districts Crosswalk**

- Directly ingested from a CSV URL.
- Used to align geographic boundaries between Parks Districts and Community Districts.

**4. NYC Dept. Homelessness Services Census**

- Pulled data via JSON API in chunks (based on row limits).
- Processed columns related to adult-only shelter populations.
- Aggregated annual averages for Community Districts.

# Processing Methodology

**Syringe Data Processing:**

- Aggregated daily syringe totals to annual totals per Parks District.
- Used crosswalk data to map Parks Districts to Community Districts.
- Divided syringe totals by number of overlapping Community Districts to estimate syringes per district.

**Controlling for Varying Geographies:**

- Addressed inconsistencies between Parks Districts and Community Districts.
- Divided Parks District syringe totals by the number of overlapping Community Districts.
- Limitations: No latitude/longitude data to enable spatial joins.

# Processing Methodology

**Housing Data Processing:**

- Imported construction data(2010–2023) for Community Districts.
- Aggregated respective five years of prior construction jobs for each syringe/ homeless collection year.
- Calculated average Year-Over-Year changes in housing jobs for 5-year window.

**Homeless Population Data Processing:**

- Summed row-specific adult shelter counts to estimate total unhoused adults for each collection date.
- Averaged annual totals by Community Districts.

**Integration:**

- Combined processed syringe, housing, and homeless population datasets for analysis.

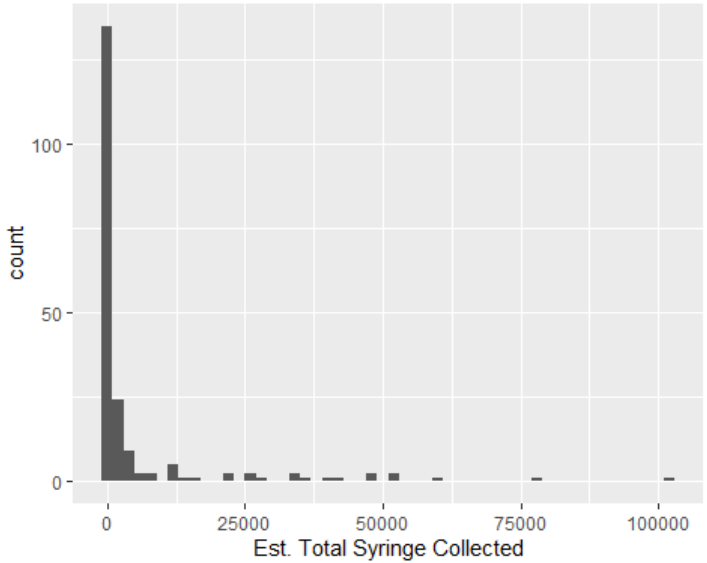
# Analysis

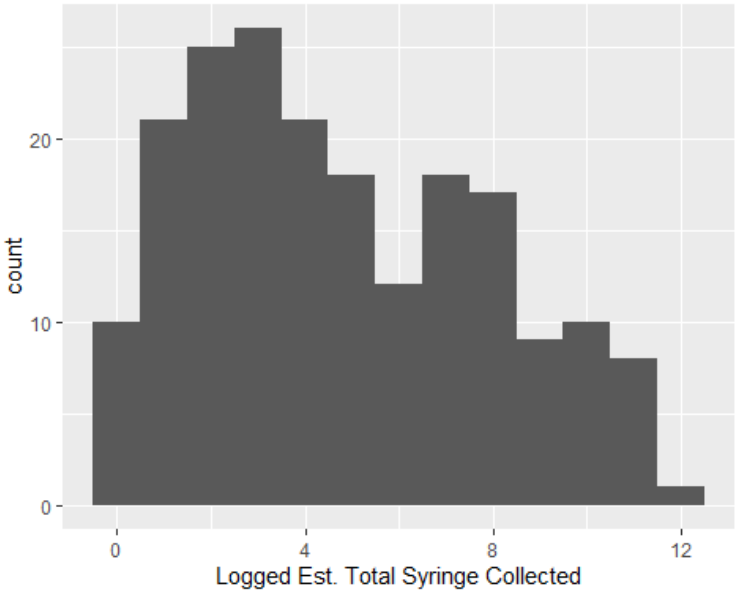
- Basic Summary Statistics

| Metric | Min | 1st Quartile | Median | Mean | 3rd Quartile | Max | NA's |
|---|---|---|---|---|---|---|---|
| Syringe Estimates | 0.5 | 7.0 | 55.0 | 4,875.9 | 1,540.8 | 10,154.7 | 374 |
| Average Homeless Count | 0.0 | 35.33 | 290.92 | 364.35 | 583.70 | 1,681.44 | 157 |
| Housing Construction (Sum) | -194.0 | 273.0 | 877.5 | 1,624.7 | 2,174.5 | 10,451.0 | |
| Housing Avg. YOY Change | -2,973.248 | 4.368 | 38.269 | 232.867 | 110.634 | 22,137.460 | 97 |

# Analysis

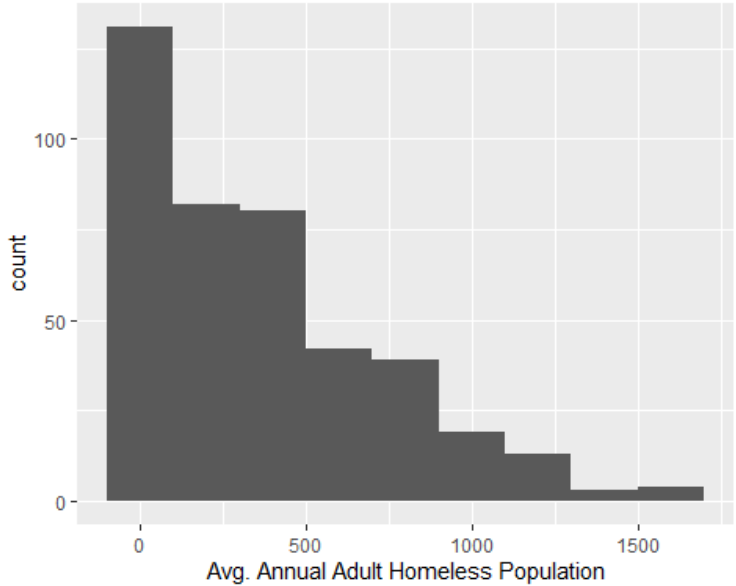
- Est. Total Syringe Counts

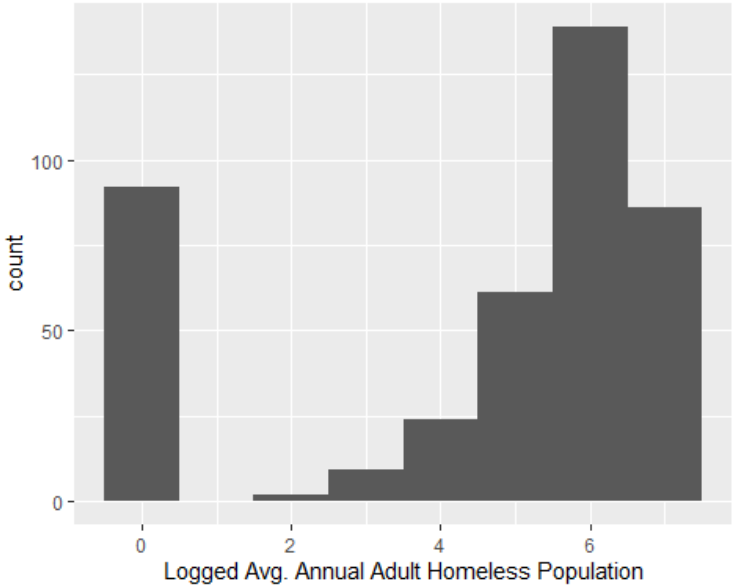


# Analysis

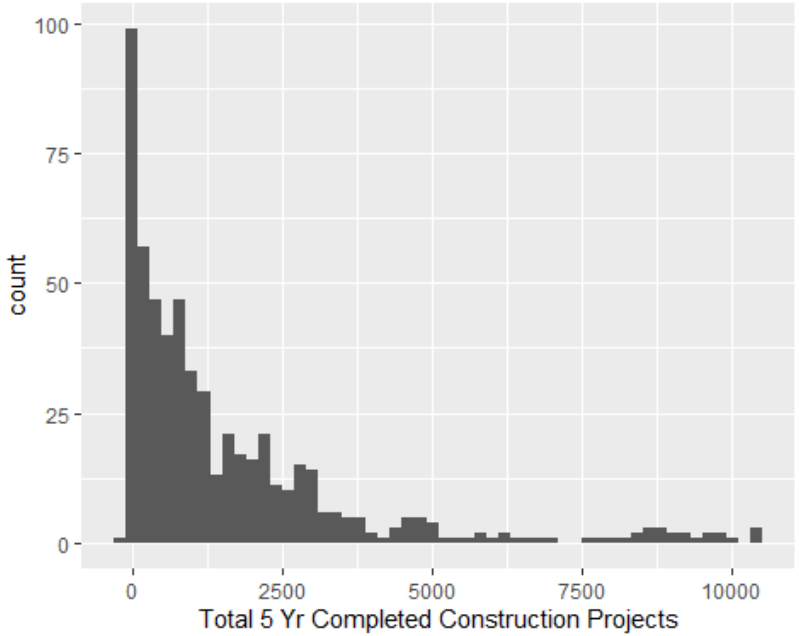
- Avg. Adult Homeless Population

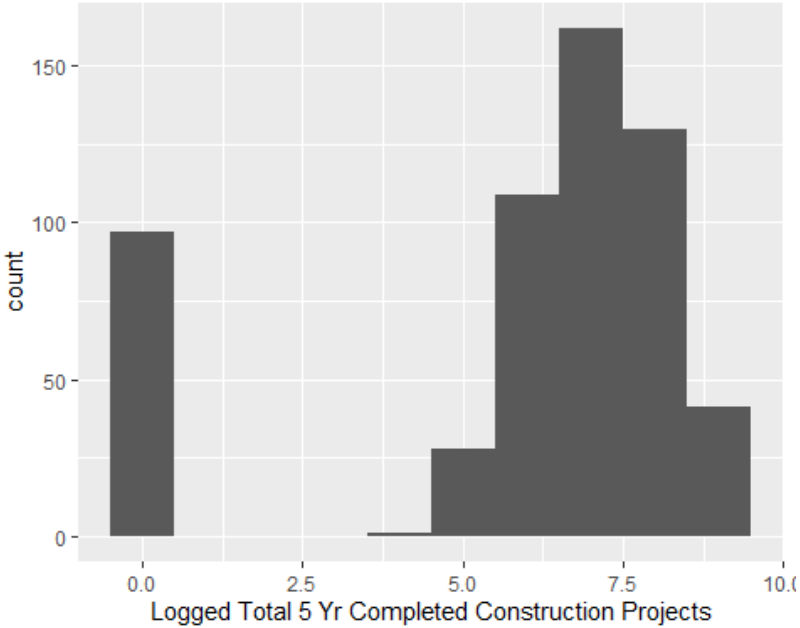


# Analysis

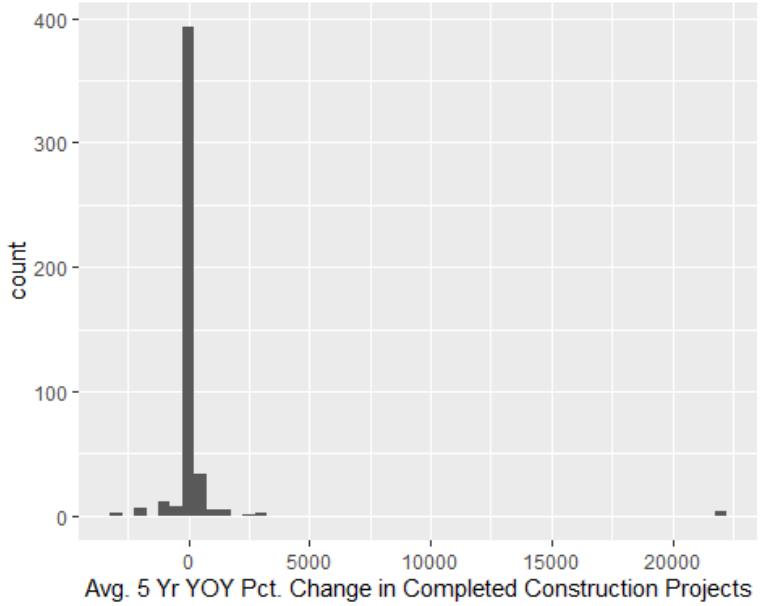
- Total Completed 5 Year Housing Related Construction Projects

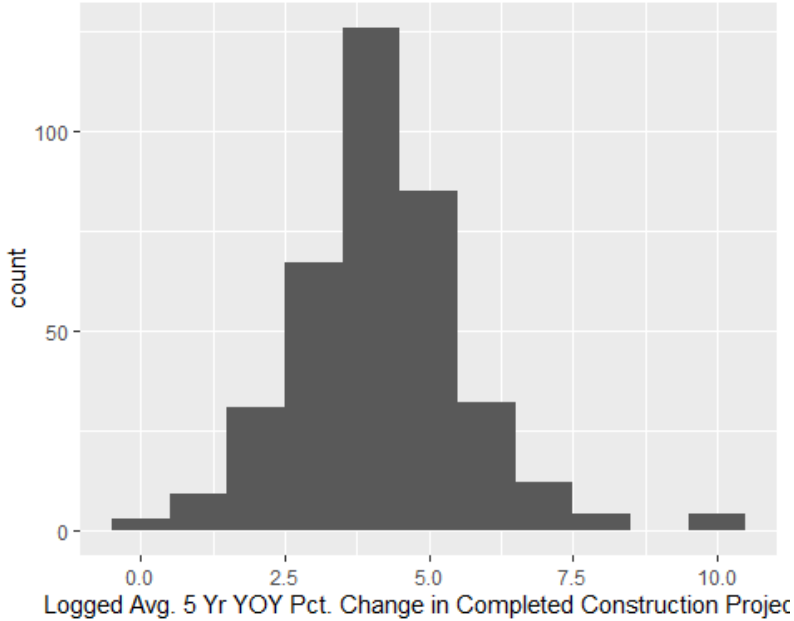


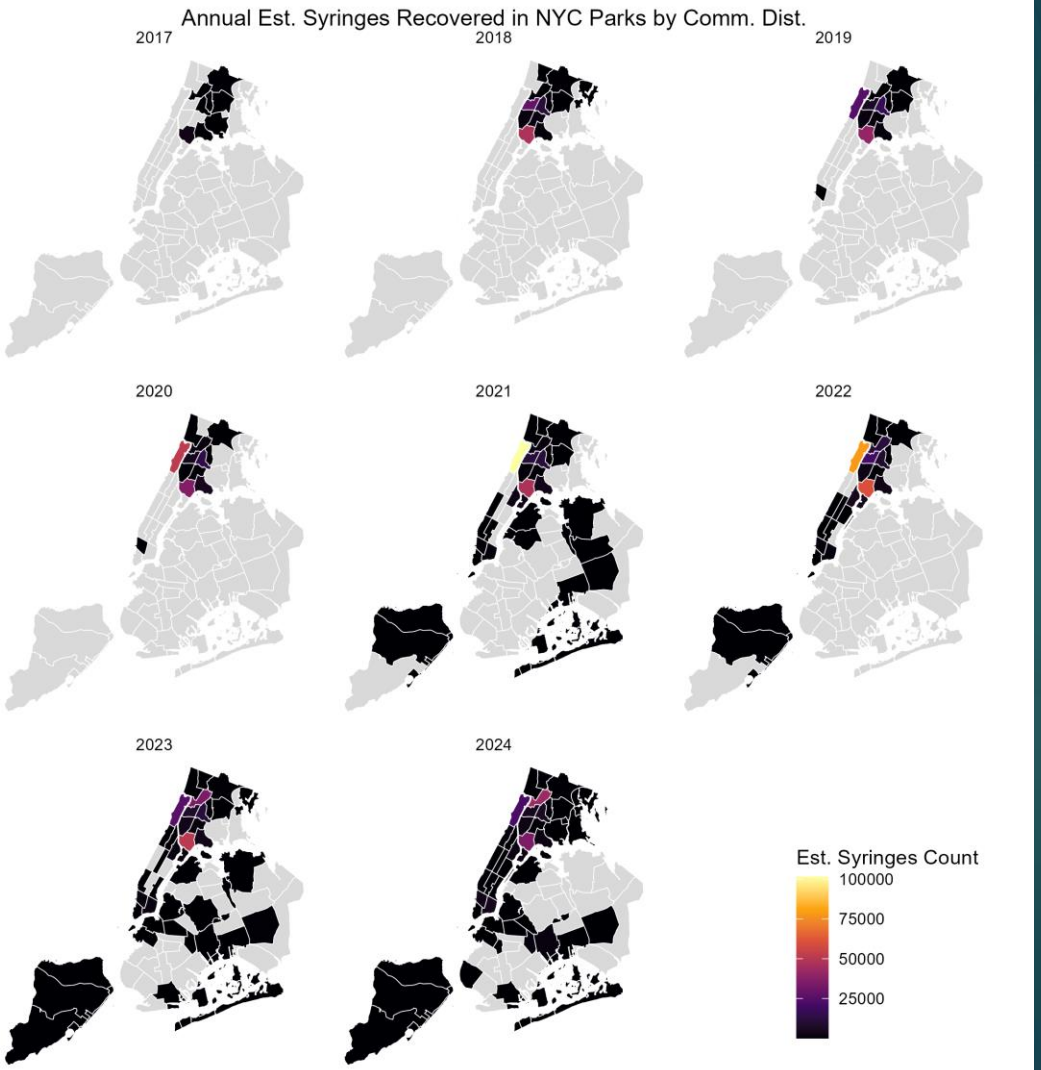



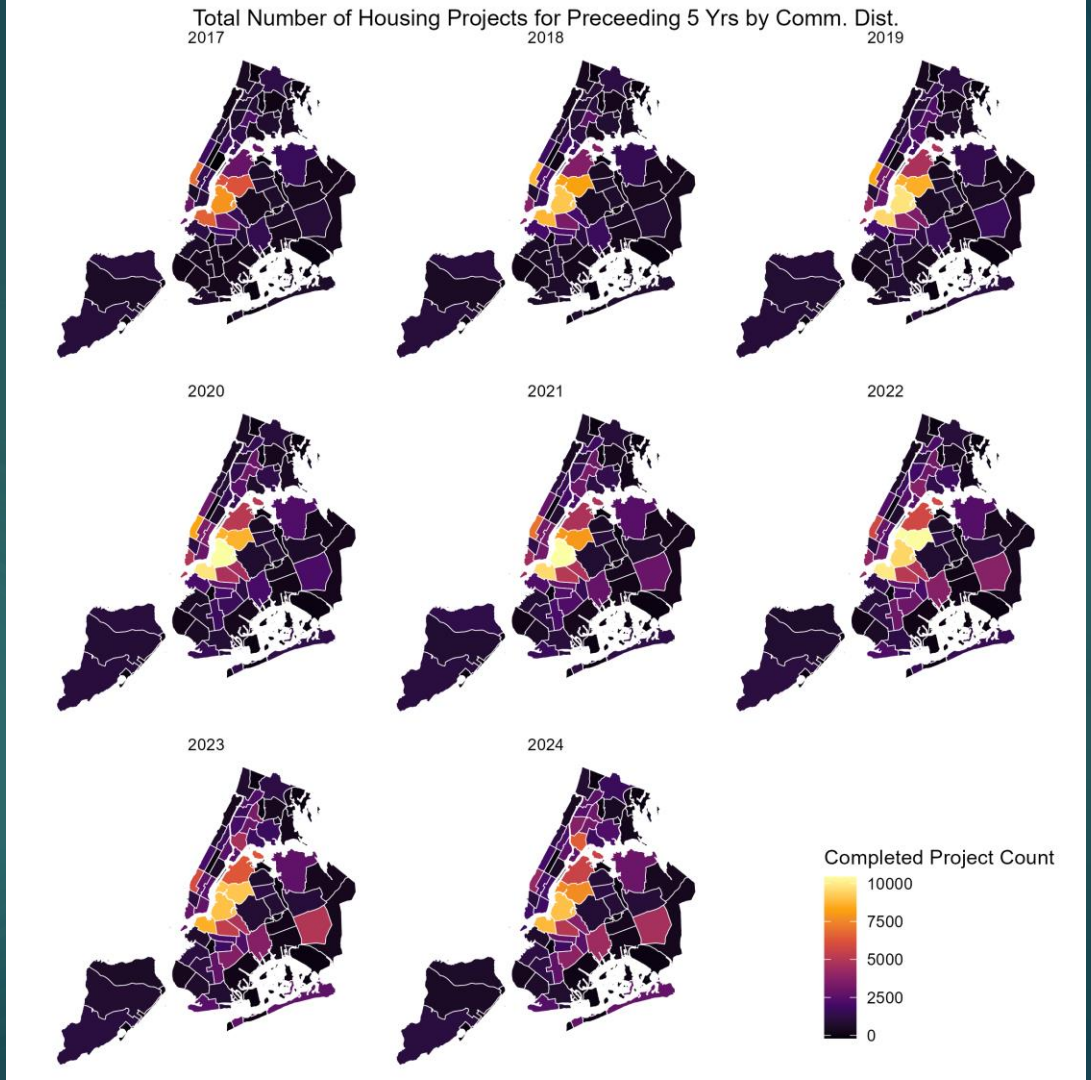
# Analysis

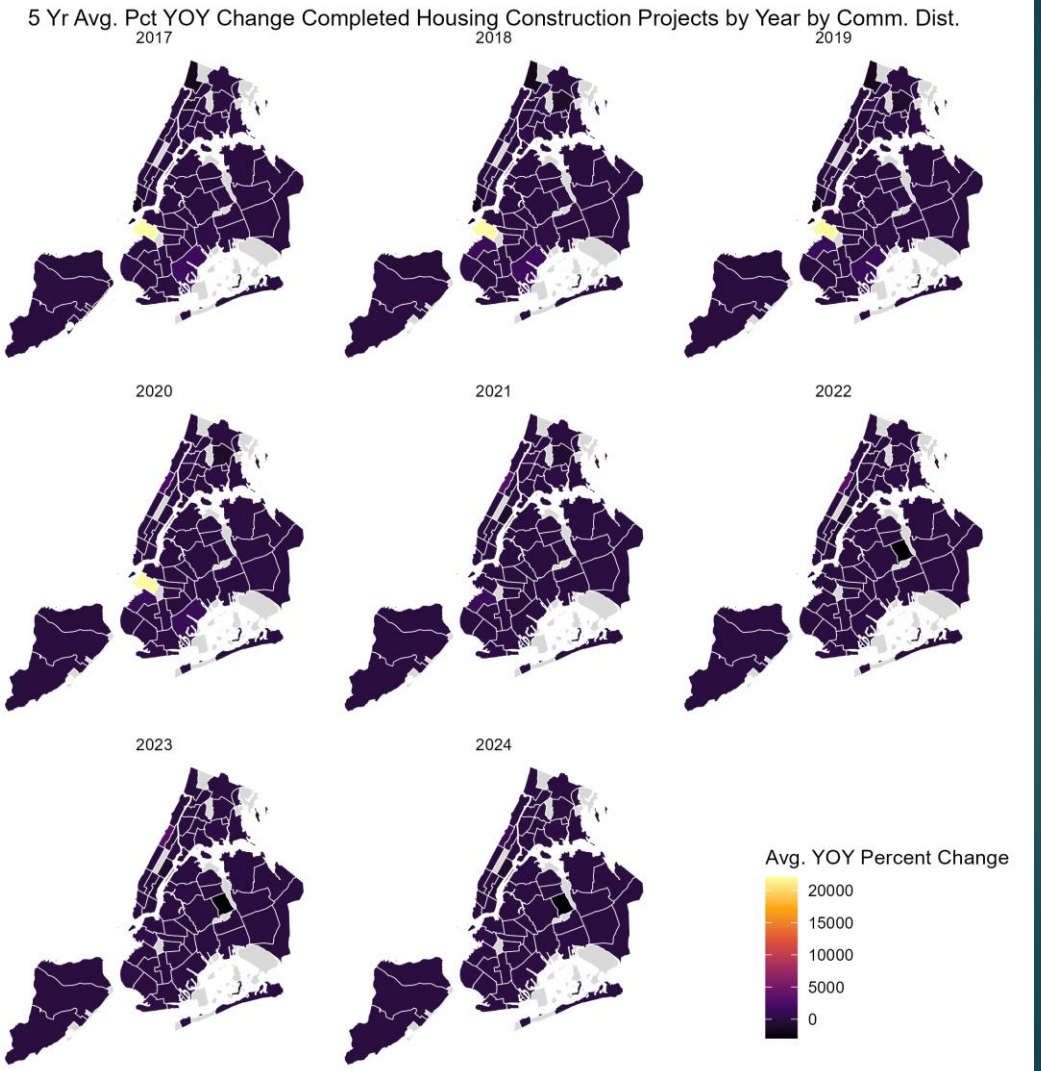
- Avg 5 Year YOY Change Housing Related Construction Projects

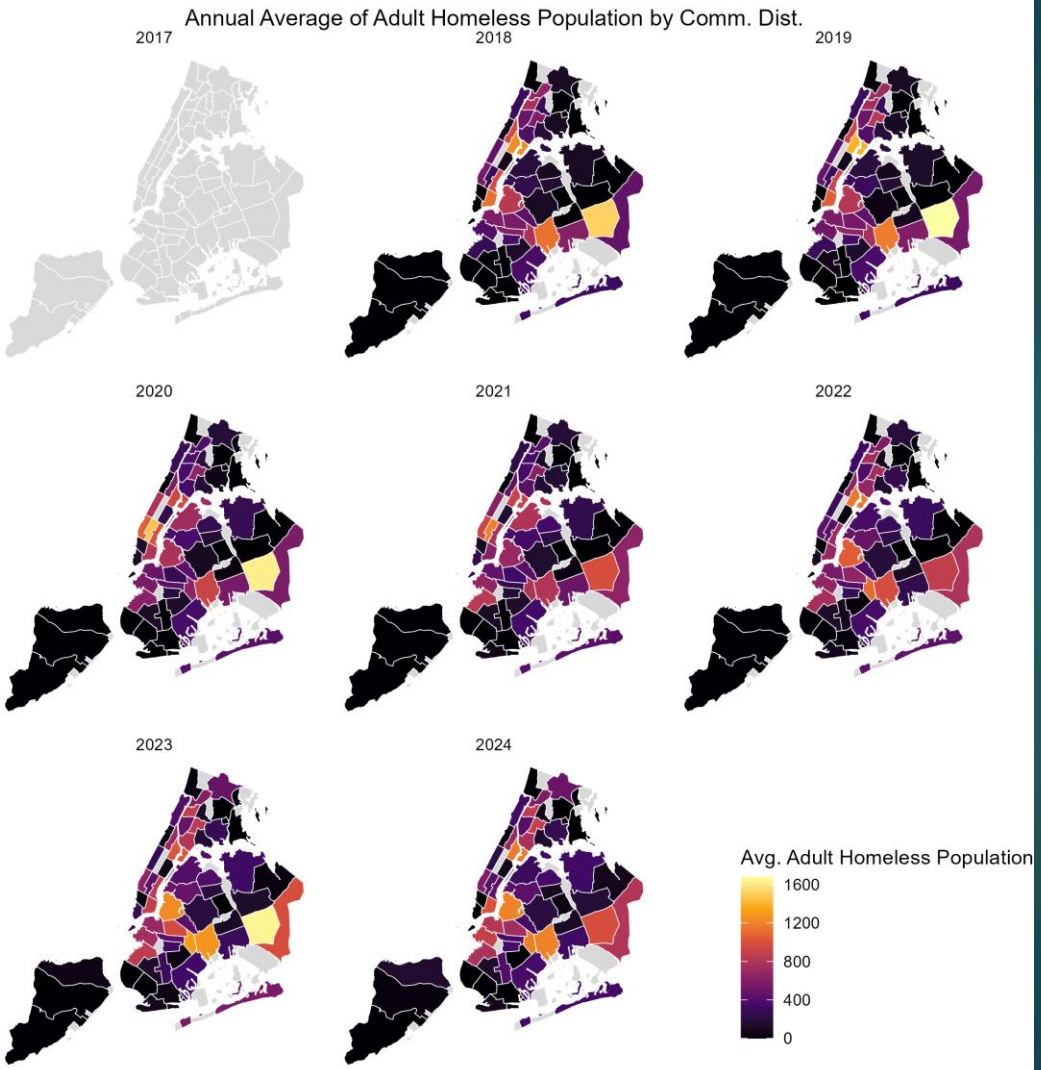



Annual Est. Syringes Recovered in NYC Parks by Comm. Dist.
2017
2018
2019
2020
2021
2022
2023
2024
Est. Syringes Count
100000
75000
50000
25000


Total Number of Housing Projects for Preceeding 5 Yrs by Comm. Dist.
2017
2018
2019
2020
2021
2022
2023
2024
Completed Project Count
10000
7500
5000
2500
0

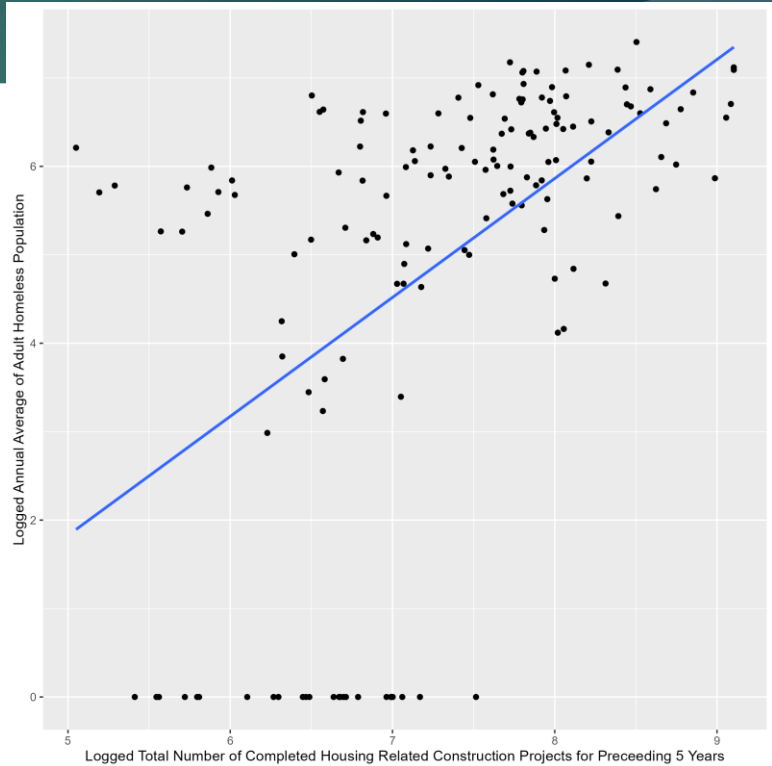

5 Yr Avg. Pct YOY Change Completed Housing Construction Projects by Year by Comm. Dist.
2017
2018
2019
2020
2021
2022
2023
2024
Avg. YOY Percent Change
20000
15000
10000
5000
0


Annual Average of Adult Homeless Population by Comm. Dist.
2017
2018
2019
2020
2021
2022
2023
2024
Avg. Adult Homeless Population
1600
1200
800
400
0

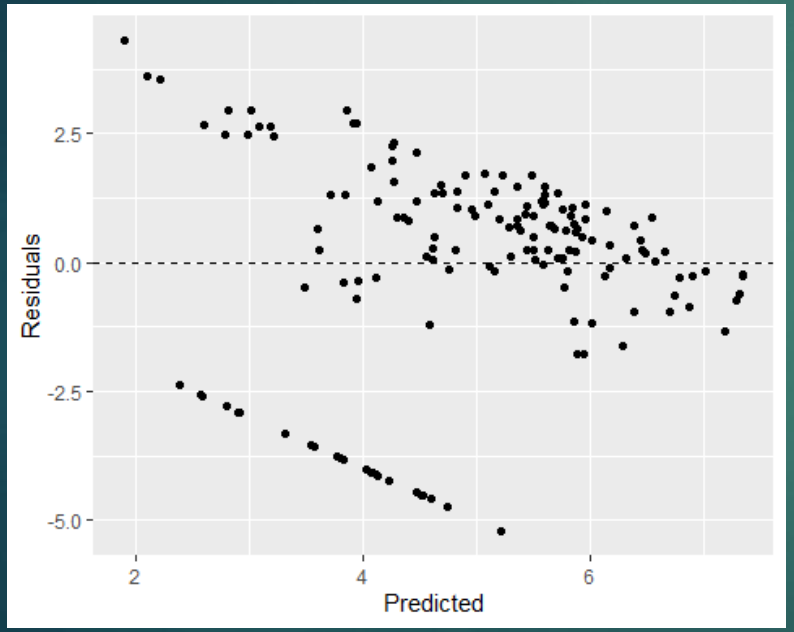
# Analysis

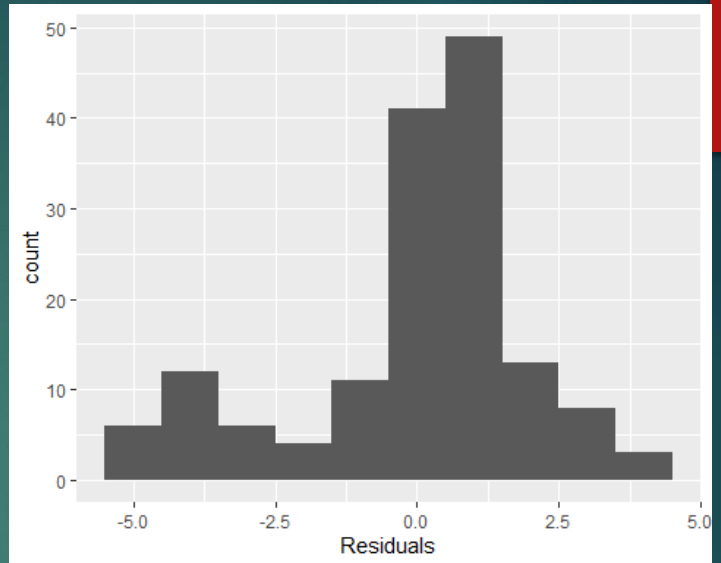
- MODEL 1: Relationship Between Logged Adult Homeless Population and Logged Completed Housing Construction Projects for Preceding 5 Years at the Community District Level

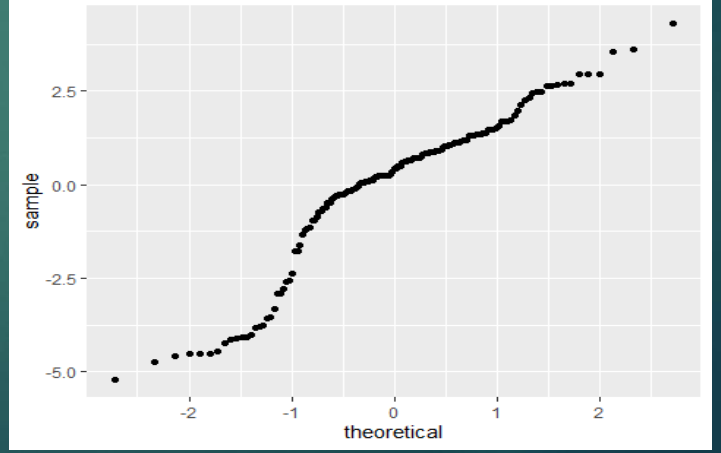
```
Call:
lm(formula = log(avg_homeless_count + 1) ~ log(housingsum + 1),
    data = data_final_nonull)

Residuals:
    Min      1Q  Median      3Q     Max
-5.2104 -0.6184  0.4339  1.1790  4.3159

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          -4.8990     1.3206   -3.71 0.000291 ***
log(housingsum + 1)   1.3453     0.1789    7.52 4.54e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.015 on 151 degrees of freedom
Multiple R-squared:  0.2725,    Adjusted R-squared:  0.2676
F-statistic: 56.55 on 1 and 151 DF,  p-value: 4.541e-12
```



# M1 Valid?


Residuals
2.5
0.0
-2.5
-5.0
2
4
6
Predicted

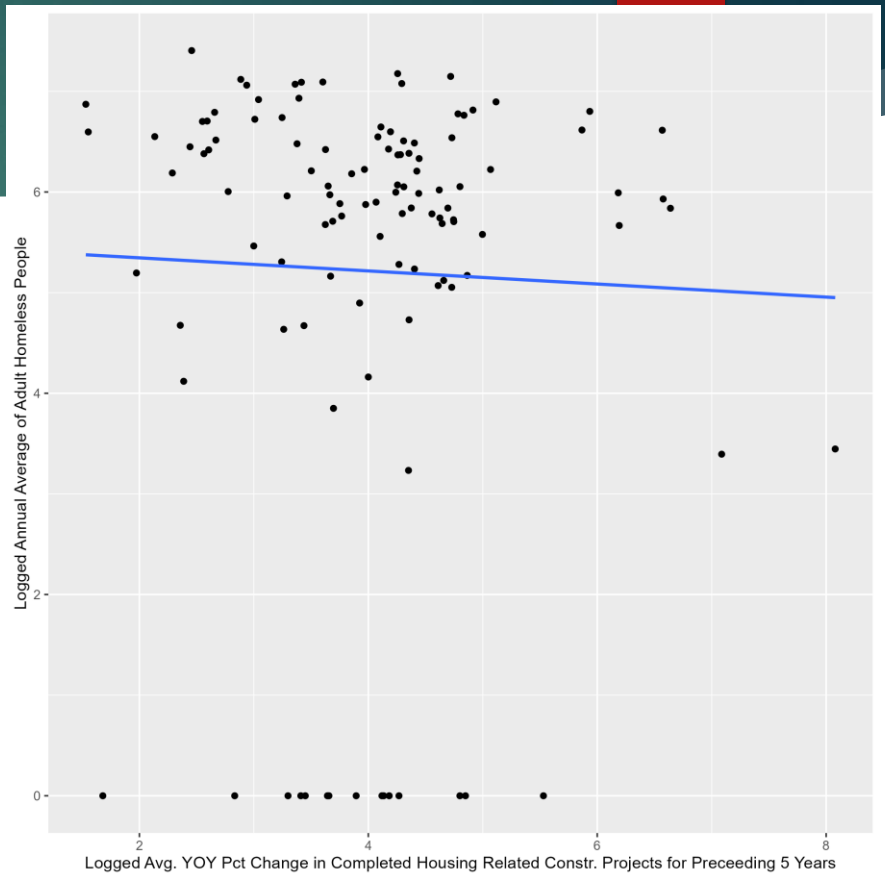

count
50
40
30
20
10
0
-5.0
-2.5
0.0
2.5
5.0
Residuals


sample
2.5
0.0
-2.5
-5.0
-2
-1
0
1
2
theoretical

# Analysis

- MODEL 2: Relationship Between Logged Adult Homeless Population and Logged Average YOY Percent Change in Completed Housing Related Construction Projects for Preceding 5 Years at the Community District Level

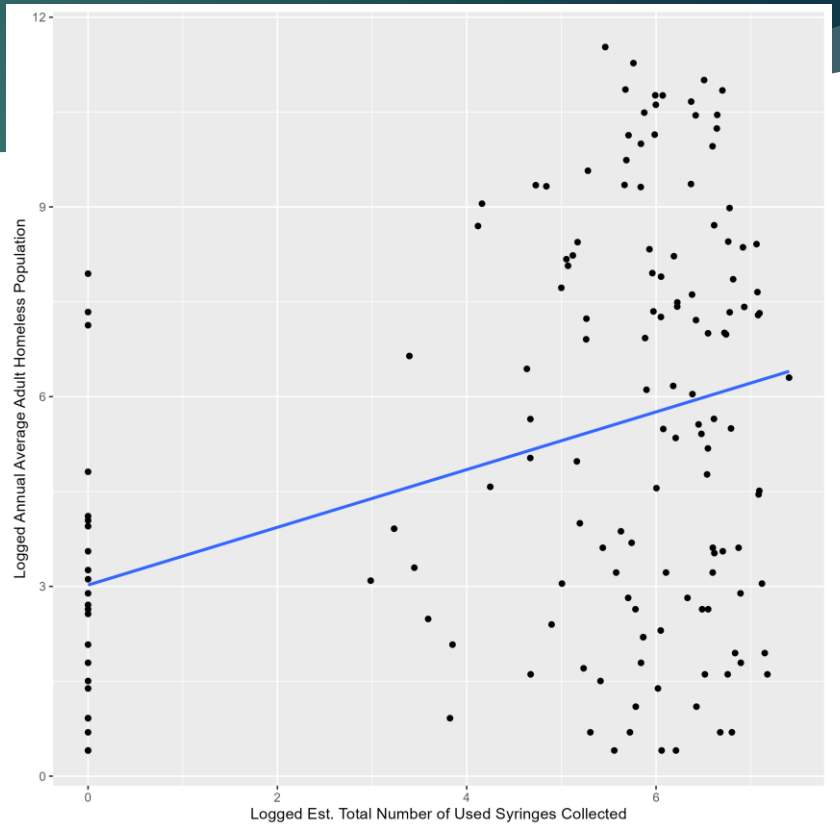
```
Call:
lm(formula = log(avg_homeless_count + 1) ~ log(prev5_avg_yoy_change +
    1), data = data_final_nonull)

Residuals:
    Min      1Q  Median      3Q     Max
-5.3665 -0.0899  0.7975  1.3047  2.0893

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                  5.47574    0.75554   7.247 5.61e-11 ***
log(prev5_avg_yoy_change + 1) -0.06501    0.18114  -0.359     0.72
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.191 on 113 degrees of freedom
  (38 observations deleted due to missingness)
Multiple R-squared:  0.001139,  Adjusted R-squared:  -0.007701
F-statistic: 0.1288 on 1 and 113 DF,  p-value: 0.7203
```



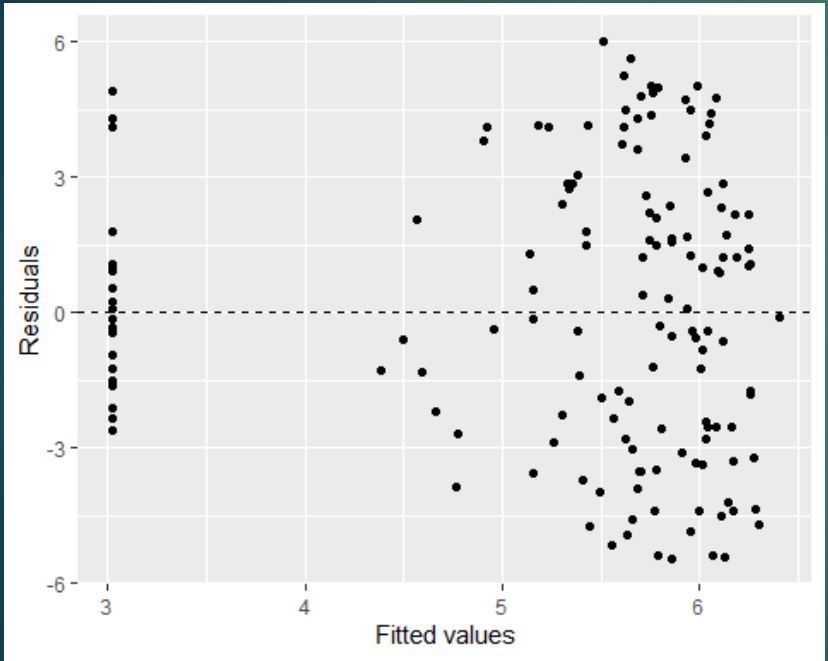
# Analysis

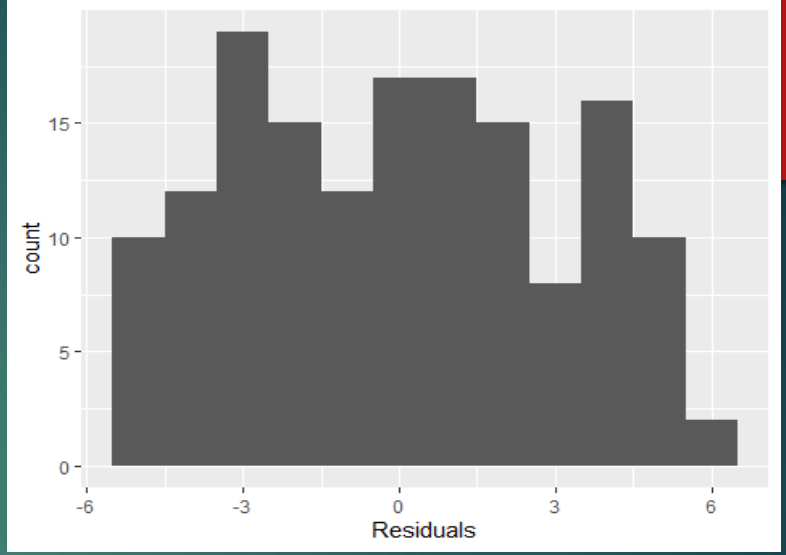
- MODEL 3: Relationship Between Logged Adult Homeless Population and Logged Est. Total Number of Syringes Recovered at the Community District Level

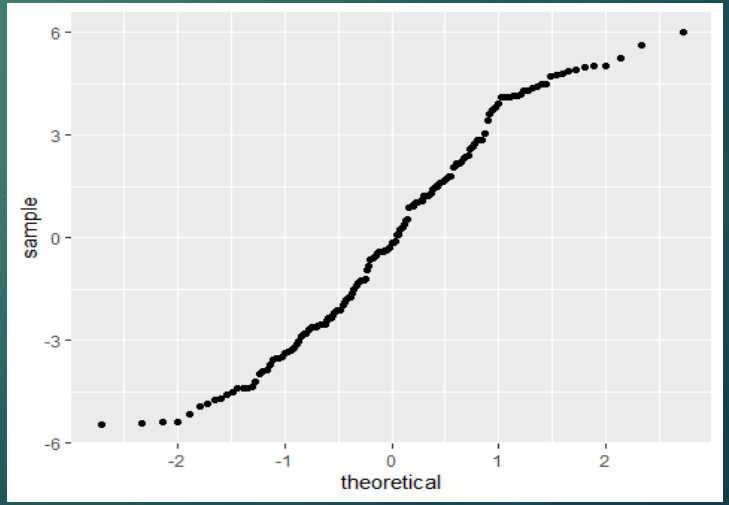
```
Call:
lm(formula = log(total_syringe_ests + 1) ~ log(avg_homeless_count +
    1), data = data_final_nonull)

Residuals:
    Min      1Q  Median      3Q     Max
-5.4495 -2.5458 -0.1314  2.3448  6.0140

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                  3.0218     0.5802   5.208 6.16e-07 ***
log(avg_homeless_count + 1)  0.4562     0.1058   4.311 2.91e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.072 on 151 degrees of freedom
Multiple R-squared:  0.1096,    Adjusted R-squared:  0.1037
F-statistic: 18.59 on 1 and 151 DF,  p-value: 2.914e-05
```



# M3 Valid?


Residuals
6
3
0
-3
-6
3
4
5
6
Fitted values


count
15
10
5
0
-6
-3
0
3
6
Residuals


sample
6
3
0
-3
-6
-2
-1
0
1
2
theoretical

# Analysis

**Challenges with Community District Data:**

- Unequal distribution of parks affecting syringe data.
- Zoning limitations influencing shelter locations and homeless population.
- Housing projects concentrated in select areas.

**Solution:**

- Aggregated data to the borough level to address CD geographic limitations.
- Hopefully reveal clearer relationships by generalizing boundaries.

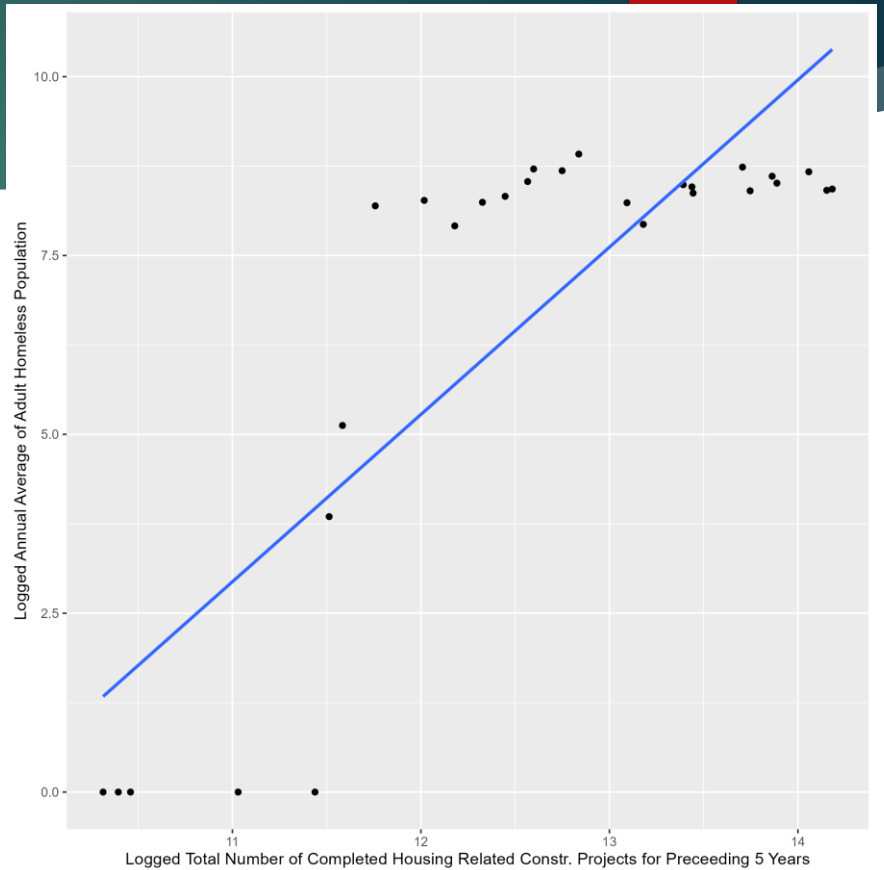
**Trade-Off:**

- Reduced data points due to aggregation.

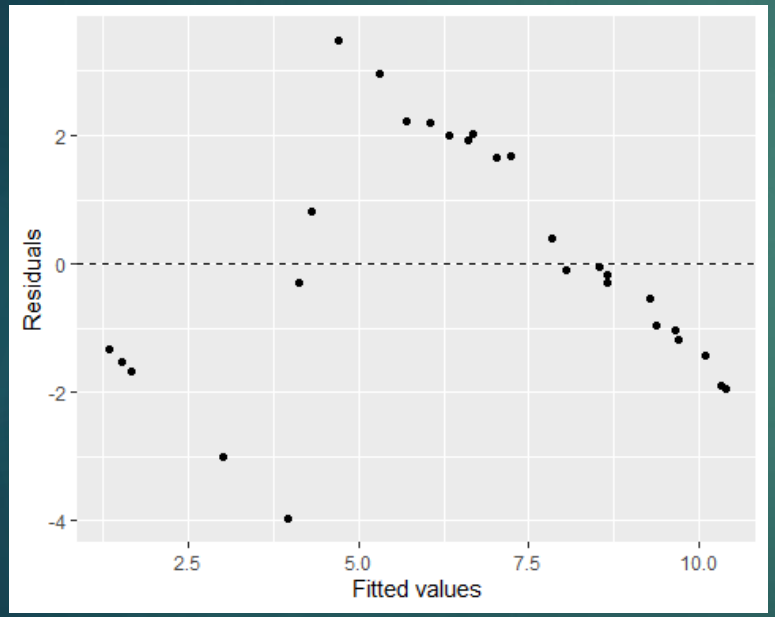
# Analysis

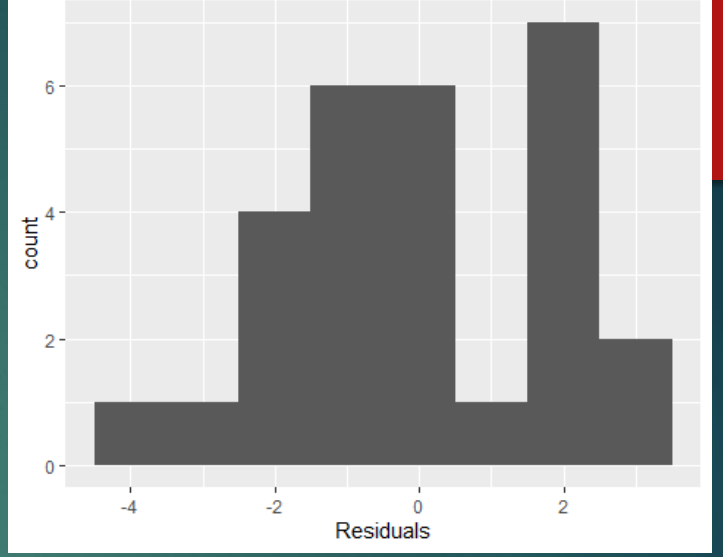
- MODEL 4: Relationship Between Logged Adult Homeless Population and Logged Total Number of Completed Housing Related Construction Projects for Preceding 5 Years at the Borough Level

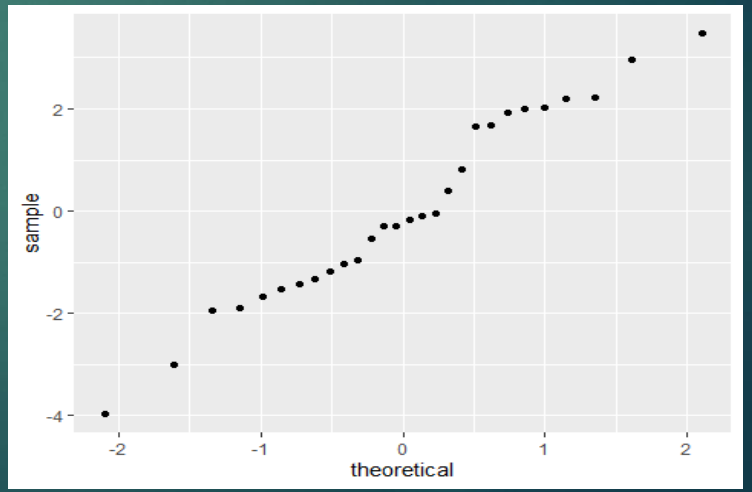
```
Call:
lm(formula = log(avg_homeless_count + 1) ~ log(housingsum + 1),
    data = data_final_boro_nonulls)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9643 -1.3569 -0.2316  1.7439  3.4821

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        -22.7823     3.9189  -5.813 3.99e-06 ***
log(housingsum + 1)   2.3384     0.3101   7.541 5.27e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.9 on 26 degrees of freedom
Multiple R-squared:  0.6862,    Adjusted R-squared:  0.6741
F-statistic: 56.86 on 1 and 26 DF,  p-value: 5.265e-08
```



# M4 Valid?


Residuals
2
0
-2
-4
2.5
5.0
7.5
10.0
Fitted values

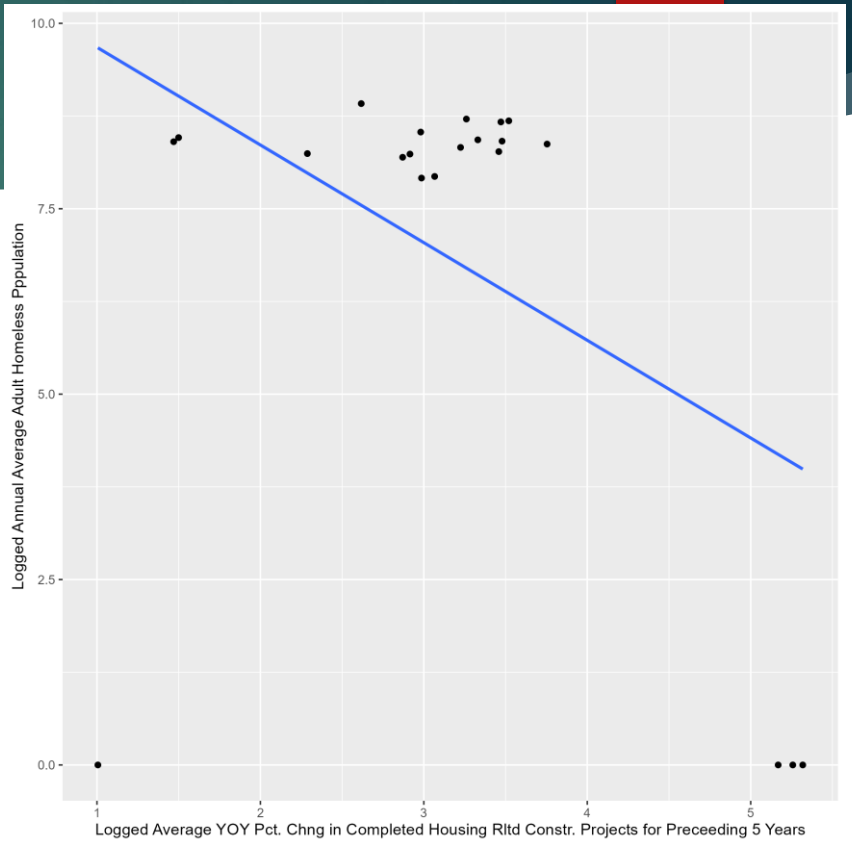

count
6
4
2
0
-4
-2
0
2
Residuals


sample
2
0
-2
-4
-2
-1
0
1
2
theoretical

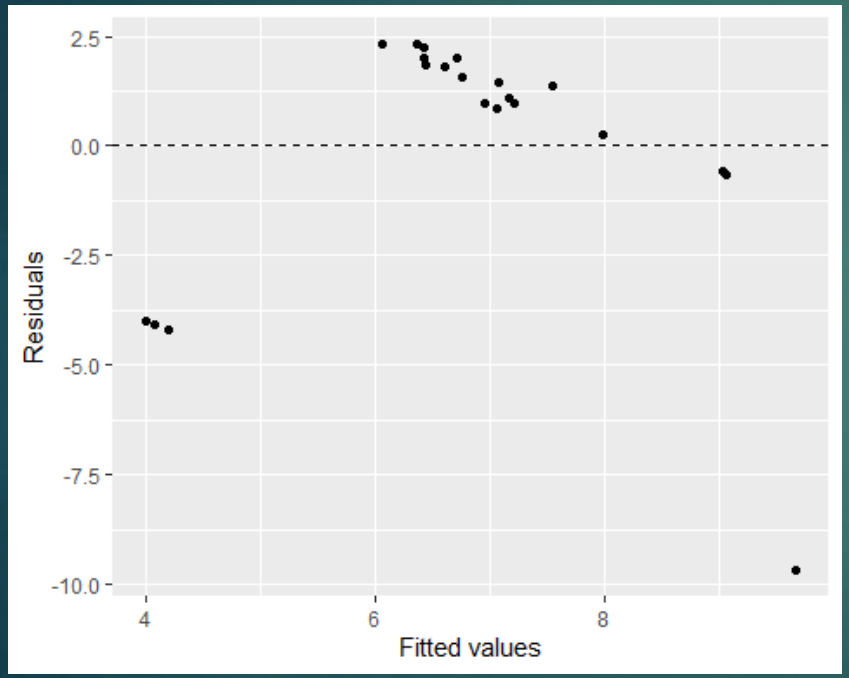
# Analysis

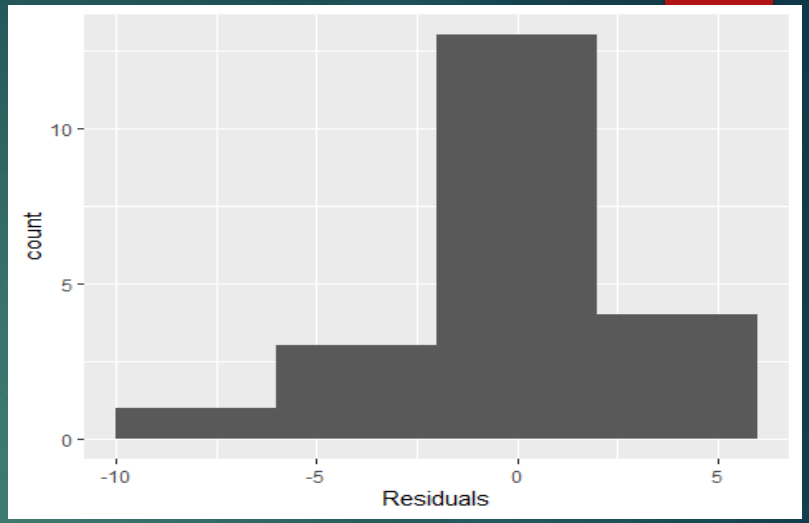
MODEL 5: Relationship Between Logged Adult Homeless Population and Logged Average YOY Percent Change in Completed Housing Related Construction Projects for Preceding 5 Years at the Borough Level

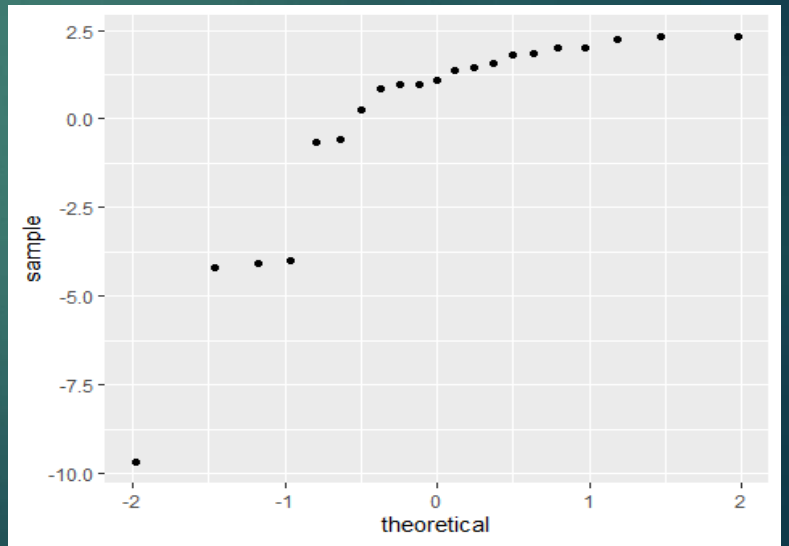
```
Call:
lm(formula = log(avg_homeless_count + 1) ~ log(avg_yoy_change +
    1), data = data_final_boro_nonulls)

Residuals:
    Min      1Q  Median      3Q     Max
-9.6693 -0.5597  1.0825  1.8315  2.3269

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            10.9937     2.0885   5.264 4.42e-05 ***
log(avg_yoy_change + 1)  -1.3169     0.6193  -2.126   0.0468 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.122 on 19 degrees of freedom
  (7 observations deleted due to missingness)
Multiple R-squared:  0.1922,    Adjusted R-squared:  0.1497
F-statistic: 4.521 on 1 and 19 DF,  p-value: 0.04681
```



# M5 Valid?


Residuals
2.5
0.0
-2.5
-5.0
-7.5
-10.0
4
6
8
Fitted values

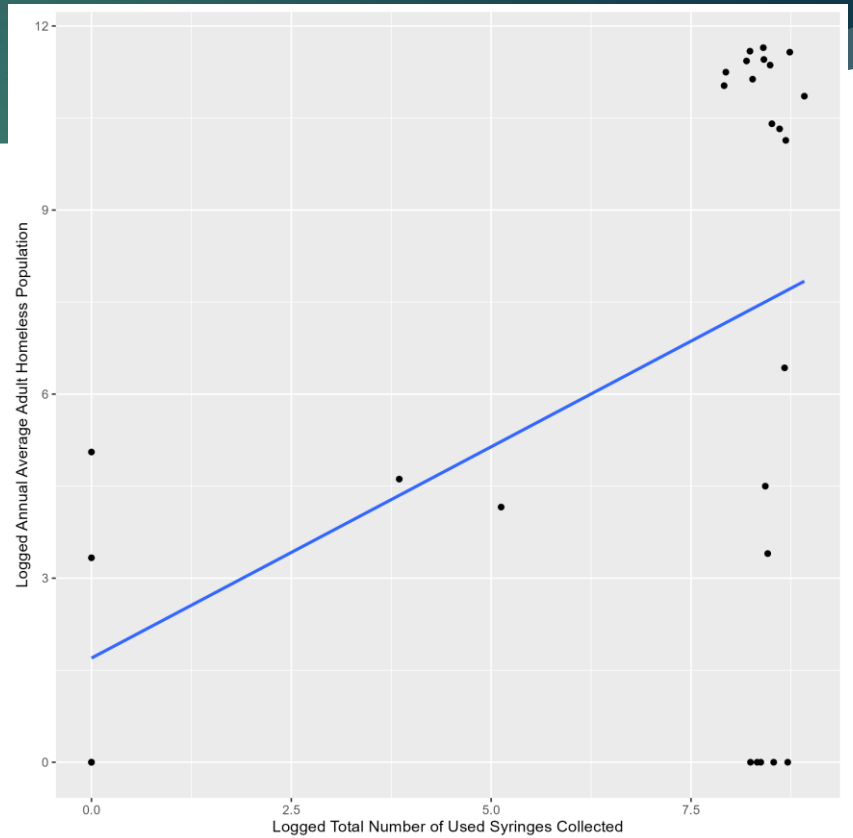

count
10
5
0
-10
-5
0
5
Residuals


sample
2.5
0.0
-2.5
-5.0
-7.5
-10.0
-2
-1
0
1
2
theoretical

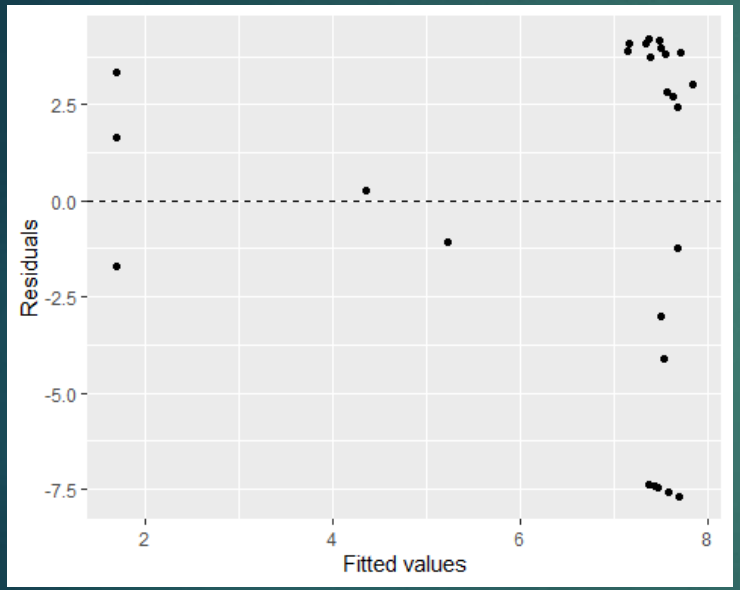
# Analysis

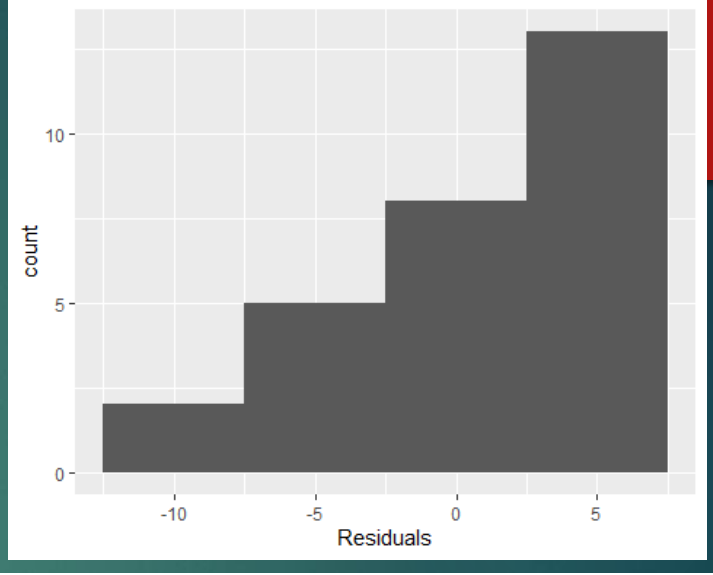
- MODEL 6: Relationship Between Logged Adult Homeless Population and Logged Total Number of Used Syringes Collected at the Borough Level

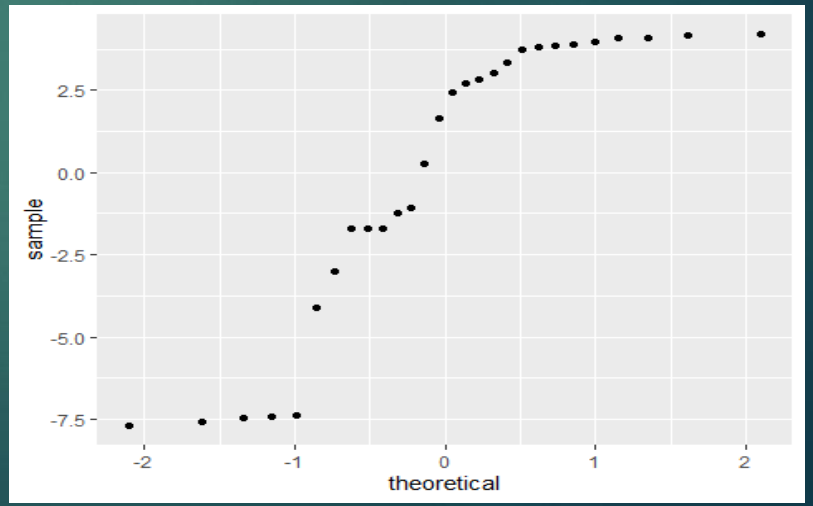
```
Call:
lm(formula = log(total_syringes + 1) ~ log(avg_homeless_count +
    1), data = data_final_boro_nonulls)

Residuals:
   Min     1Q Median     3Q    Max
-7.697 -2.024  2.045  3.827  4.218

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                  1.6974     1.8896   0.898    0.377
log(avg_homeless_count + 1)  0.6889     0.2552   2.699    0.012 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.414 on 26 degrees of freedom
Multiple R-squared:  0.2189,    Adjusted R-squared:  0.1889
F-statistic: 7.287 on 1 and 26 DF,  p-value: 0.01205
```



# M6 Valid?


Residuals
2.5
0.0
-2.5
-5.0
-7.5
2
4
6
8
Fitted values


count
10
5
0
-10
-5
0
5
Residuals


sample
2.5
0.0
-2.5
-5.0
-7.5
-2
-1
0
1
2
theoretical

# Statistically Significant Regression Results

| Regression Analysis | Geography Level | Dependent Variable | Independent Variable | R-squared | Adjusted R-squared | P-value | Correlation Type | Validity |
|---|---|---|---|---|---|---|---|---|
| Adult Homeless Pop. & Total Housing Proj. Prev 5 Yrs. (Logged) | Community District | log(avg_homeless_count + 1) | log(housingsum + 1) | 0.2725 | 0.2676 | 4.54e-12 | Direct | May not be valid |
| Adult Homeless Pop. & Total Recovered Syringe Count (Logged) | Community District | log(total_syringe_ests + 1) | log(avg_homeless_count + 1) | 0.1096 | 0.1037 | 2.914e-05 | Direct | Should be Valid |
| Adult Homeless Pop. & Total Housing Proj. Prev 5 Yrs. (Logged) | Borough | log(avg_homeless_count + 1) | log(housingsum + 1) | 0.6862 | 0.6741 | 5.27e-08 | Direct | May not be valid |
| Adult Homeless Pop. & Avg. YOY Pct. Chg. Housing Proj. Prev 5 Yrs. (Logged) | Borough | log(avg_homeless_count + 1) | log(avg_yoy_change + 1) | 0.1922 | 0.1497 | 0.04681 | Indirect | May not be valid |
| Adult Homeless Pop. & Total Recovered Syringe Count (Logged) | Borough | log(total_syringes + 1) | log(avg_homeless_count + 1) | 0.2189 | 0.1889 | 0.01205 | Direct | May not be valid |

# Results and Conclusion

**Community District Level Results**

Overall, the validity of the models is in question, they may not be valid.

**Model 1:** Direct relationship between completed housing projects and adult homeless population.

- **Key Stats:**
  - 27% of variance in homeless population explained by housing projects.
  - Unexpected result: Homeless population increased as housing projects increased.
- Rejects null hypothesis; confirms alternative hypothesis.

**Model 2 :** Not Statistically Significant

**Model 3:** Direct relationship between syringes collected and adult homeless population.

- **Key Stats:**
  - 10% of variance in syringe counts explained by homeless population size.
- Rejects null hypothesis; confirms alternative hypothesis.

# Results and Conclusion

**Borough Level Results**

Overall, the validity of the models is in question, they may not be valid.

**Model 4:** Direct relationship between housing projects and homeless population.

- **Key Stats:**
  - 67% of variance in homeless population explained by housing projects.
  - Stronger correlation compared to community district level.

**Model 5:** Indirect relationship between YOY change in housing projects and homeless population.

- **Key Stats:**
  - 15% of variance in homeless population explained by YOY housing change.
  - Larger YOY increases correlate with a decrease in homeless population.

**Model 6:** Direct relationship between syringes collected and homeless population.

- **Key Stats:**
  - 19% of variance in syringe counts explained by homeless population.

# Results and Conclusion

**Takeaways**

- An unexpected direct relationship exists between completed housing projects and homelessness. This may mean that many high-end developments in wealthier, gentrifying areas may be correlated a potential exacerbation of affordable housing shortages rather than alleviating homelessness.
- A Year-Over-Year (YOY) increase in housing projects correlates with reduced homelessness. This may indicate efforts to address housing needs in underserved areas, where development helps mitigate the issue.
- The direct relationship between homelessness and syringe recovery underscores how reducing homelessness could improve public health outcomes in this dimension.

# Results and Conclusion

**Issues for Next Steps**

- The data used lacks differentiation between luxury developments and affordable housing. Key metrics like total units added, affordable units created, or rent-stabilized units removed are absent. More granular data is needed.
- Differences in zoning, the presence of parks, and shelter access across Community Districts/ Boroughs may have skewed numbers. City wide analysis over more years, or targeted case studies may be more ideal or this type of analysis.
- Shelter-based data excludes unsheltered individuals, leaving a significant gap in understanding the true scale of homelessness.