Final Project - DATA 606 & DATA 607

John Ferrara

2024-12-08

Abstract

This observational study examines whether completed housing-related construction correlates with the adult homeless population, and how homelessness correlates with public syringe recovery in NYC parks. Using data sourced from NYC Open Data, two questions were analyzed: (1) Does completed housing construction projects over the preceding five-year window affect the size of the adult homeless population? (2) Does the size of the adult homeless population correlate with disposed syringe recovery rates? Data from the NYC Department of City Planning for construction projects, NYC Parks Department syringe recovery data, and NYC Department of Homelessness Services homeless population data were analyzed using linear regression models, both at the community district and borough levels.

Several statistically significant relationships were found. At the community district level, more completed construction projects directly correlated with a larger homeless population, possibly reflecting the concentration of construction jobs in wealthier or gentrifying areas. Additionally, an increase in the adult homeless population directly correlated with higher syringe recovery rates in NYC parks. At the more generalized borough level, the correlation between completed construction projects and homelessness was stronger, with the construction projects explaining roughly 67% of the variance in homelessness populations. Borough-level analysis also revealed an indirect relationship between the average year-over-year increases in completed construction projects and homelessness, suggesting targeted development may help reduce homelessness.

These findings highlight connections between housing policy, homelessness, and syringe recovery as a singular dimension of public health. While limitations for this analysis exist, such as a lack of housing data granularity and a lack of non-shelter based homeless counts, the analysis underscores the importance of public data in addressing complex urban challenges.

Introduction and Overview

As most know, in recent years the cost of housing has been outpacing wage growth1. Rising housing costs have led to a series of policy issues. While some impacts of high housing costs are obvious, others are less apparent. This analysis explores whether one such issue — the recovery of used syringes in public parks — is correlated to the housing crisis. In short, the overarching question: Does housing impact this seemingly separate public health and safety issue of public, or improper, syringe disposal?

The overarching question for this observational study must be broken down into two different main questions. The following are the two sub-questions and their respective hypotheses:

- *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.
  - *Independent Variable*: 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 Variable*: 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.
  - *Independent Variable*: The independent variable in this question is the adult unhoused population.
  - *Dependent Variable*: The dependent variable in this question is the total syringes collected from NYC Parks and public safe disposal sites.

It should be noted that these are large questions. Questions that are much more nuanced and complex than any single analysis can outline, so while this study seeks to demonstrate a correlation between varying dimensions of public policy implications, it is by no means an attempt at a solution. Rather, the study seeks to determine if there is an overlap between these issues using several sources of public data.

Data Sources

There are multiple data sources used in this study. Most, if not all, were sourced from NYC Open Data (https://opendata.cityofnewyork.us/) a website that provides a multitude of public data sets generated by New York City government agencies. The data used in this analysis were:

- Dataset 1: NYC Parks Syringe Collection Data (https://data.cityofnewyork.us/Public-Safety/Summary-of-Syringe-Data-in-NYC-Parks/t8xi-d5wb/about_data)

Overview & Data Assumptions

This dataset is from the NYC Parks Department. NYC Parks department staff, along with the staff of various community non-profit organizations, collect used syringes discarded improperly in public parks and log the totals. These collection totals are grouped by each Parks District, which is an internal administrative geographic boundary for the New York City Parks Department. Each row of data is the equivalent of a day's total collection of syringes. Within this data set there are three total columns that outline the number of syringes collected. They are as follows:

- Total Kiosk Syringes Collected: This is the total number of syringes that are collected from the city's safe disposal kiosks.
- Total Ground Syringes Collected: This is the total number of syringes that are collected from the ground, or just generally found to be improperly disposed of.
- Total Syringes Collected: This is the total number of syringes from both the "ground" category and the "kiosk" category.

This analysis makes use of the total ground syringes collected, not the kiosk or ground syringes. While assumptions could be made about housed and unhoused population's syringe disposal habits, the total number of syringes publicly disposed by either method should capture what is needed for this analysis.

Source Format & Ingestion Method

Sourced from CSV formatted files. The NYC Open Data API was used for iteratively ingesting this data from CSV structured formatting.

- Dataset 2: NYC Dept. City Planning Housing Database (https://data.cityofnewyork.us/Housing-Development/Housing-Database/6umk-irkx/about_data)

Overview & Data Assumptions

This data is from the NYC Department of City Planning (DCP), the data contains information from the NYC Department of Buildings (DOB) for how many construction and demolition jobs are within a specific municipality-based geographies. For this analysis used at Community Districts for the main geographic boundary. The analysis leverages yearly counts of completed construction projects. The annual count data does not provide nuance between the three main types of construction projects included in the data (new buildings, major alterations, and demolitions), however, for the sake of this analysis, a larger number of projects is assumed to be targeted at enhancing living conditions and making new units available to the housing market. Annual counts from 2010 through 2020 by Community District were used in varying relative 5-year windows. How this was done is explained in detail within the methodology section.

The original download is a zipfile with multiple geographies, for this analysis the Community District boundary was used.

Source Format & Ingestion Method

The data source was downloaded in a zipfile containing multiple csv files for various geographic boundaries, the data was extracted, and a singular csv file containing housing units by Community District was ingested for processing.

- Dataset 3: NYC Parks Disticts (https://data.cityofnewyork.us/City-Government/NYC-Parks-Districts/mebz-ditc/about_data)

Overview & Data Assumptions

This data set contained both the NYC Parks Department districts and Community Districts within one file. This geography data was processed so as to flatten the number of Community Districts that were listed for each respective Parks District. For example, if one row of data for Parks District X listed overlaps with Community Districts A, B and C. These three districts, in the raw data, are listed in one cell. This data was parsed so that the one row was extracted into three different rows, having one row for each Community District associated with the Parks District. The data was essentially flattened long ways, the data was subsequently grouped to get a unique value count for Community Districts per Parks District. The processed table was used as a crosswalk to process the syringe numbers to obtain an estimate of syringe counts at the Community District level. As mentioned, the NYC Syringe Collection data only contained

Parks districts, this crosswalk was joined into the syringe data in order to get syringe count estimates for Community Districts. This is discussed further in the methodology section.

Source Format & Ingestion Method

Read in directly from a single CSV URL.

- Dataset 4: NYC Dept. Homelessness Services Individual Census by Borough, Community District, and Facility Type (https://data.cityofnewyork.us/Social-Services/Individual-Census-by-Borough-Community-District-an/veav-vj3r/about_data)

Overview & Data Assumptions

This dataset contains counts of individuals within the various types of shelters across New York City by community district for various reporting dates. The shelter types covered by this dataset include:

- Adult Family (Commercial Hotel): The population of adults in makeshift homeless shelters held within commercial hotels designated for adult families (i.e.m married couples with no children, a family with no children under the age of 21, or an unmarried couple who meets the DHS definition of a family unit).
- Adult Family Shelter: The population of adults in homeless shelters designated for adult families (i.e.m married couples with no children, a family with no children under the age of 21, or an unmarried couple who meets the DHS definition of a family unit).
- Adult Shelter: The population of adults in homeless shelters designated for single unhoused adults.
- Adult Shelter (Commercial Hotel): The population of adults in makeshift homeless shelters held within commercial hotels designated for single unhoused adults.
- Family with Children (Commercial Hotel): The population of individuals in makeshift homeless shelters held within commercial hotels designated for families with children.
- Family with Children Shelter: The population of individuals in homeless shelters designated for families with children.

For the sake of this analysis, only those shelters categories with adult-only numbers were counted, the assumption being is that intravenous drug users would be adults and not families with children. This means that the totals for Adult Family (Commercial Hotel), Adult Family Shelter, Adult Shelter, and Adult Shelter (Commercial Hotel) were used in the analysis for homeless populations within a community district. Those columns that had counts for family-only specific counts were not used.

Source Format & Ingestion Method

Sourced from JSON formatted files. The NYC Open Data API was used for iteratively ingesting this data from JSON structured formatting.

Methodology

After ingesting the various data sets through their respective means, the data was processed. Each data set needed a unique series of processing steps in order to yield the finalized version of the data ultimately included in this analysis.

Syringe Data Processing

The Syringe data had totals for each park district on each date that the syringes were collected. For the sake of this analysis, the numbers were aggregated up to an annual collection total for each park district. Using the processed Community District to Park District crosswalk data, the final version of the crosswalk contained Community District, Parks District, and a count of distinct Community Districts for each parks district. The processed versions of both dfs were joined, and in order to yield an estimate for ground syringes recovered the Park District Syringe totals, which were native to the data, were divided by the number of community districts overlapping with each Parks District to generate a syringe estimate for each community district.

Controlling for Varying Geographies

As mentioned, the incongruities of the geographic units between the Community District-based housing and homeless datasets and the NYC Parks District syringe collection data set mandated the raw syringe totals be processed to yield a syringe estimate for Community District. Community Districts are a city-wide administrative boundary associated with the city’s community boards, while the parks districts are an internal NYC Parks Department administrative boundary. These two geographic boundaries do not have a 1:1 relationship with each other. There were instances of a singular community district overlapping with multiple parks districts, and vice versa. In order to approximate the number of syringes per community district, I simply divided the total number of syringes for each parks drastic by the total number of distinct Community Districts in each respective parks district. While this methodology is imperfect, the syringe collection data did not have Latitude and Longitude for each collected syringe or for specific collection sites, so a spatial join using shapefile geographies could not be completed.

Housing Data

The housing data that was used was imported after downloading a zip file from 2023 Q4, and using one Community District-specific CSV from the batch of files. The CSV used contains annual counts fr the number of completed housing constructions projects fr each year from 2010 through 2023 by community district. For this analysis the assumption was made that the number of syringes collected in an area, provided that syringe collection is impacted by new housing units, would be a lagging indicator. With this mindset, the number of completed housing construction projects in a community district for five years prior to the year the syringes were collected, was the aggregate version of housing numbers used for the analysis In other words, for each year in the syringe collection data, the preceding five years of construction jobs were summed up and the average Year-Over-Year change in construction jobs for those years were calculated. For instance, for the total number of estimated syringes collected for a Community District in 2017, the metrics for housing-related construction jobs within that same community district would be the aggregate sum of all housing jobs from 2011 through 2016, as well as the average Year-Over-Year change in housing jobs for those years. This aggregation into two main columns for each annually summed community district row allowed for the housing numbers to be joined with the Syringe Collection data.

Homeless Population Data

The Homeless shelter population data was pulled into the analysis in JSON format via NYC Open Data’s API. To process this data the sum of the following columns were used as an estimate for the total number of unhoused adults: adult_shelter, adult_shelter_comm_hotel, adult_family_shelter,and adult_family_comm_hotel. These columns were counts from various types of shelters throughout the city. The row-specific sums for these columns were then averaged for each community district on an annual basis, which was to control for multiple collections within a year. Once completed, the processed data was joined into the df that contained the syringes data and the housing construction data.

Ingestion

NYC Parks Syringe Collection Data

```
## API Intake; Documentation states there are 35,155 rows of data. Iterate through these numbers for API intake ( L
imited to 1,000 row chunks)
total_srg_rows <- 35155 #Manually Checked on Site
srg_endpnt <- "https://data.cityofnewyork.us/resource/t8xi-d5wb.csv"
srg_suffix <- '?$offset='

##Initial Pull
initial_pull <- read.table(srg_endpnt, header = TRUE, sep = ",", dec = ".")
offset <-nrow(initial_pull) # Should be the length of the API chunk limit. (e.g., 1000 rows)

##Iteratting through all of the chunks & Appending to empty DF
running_pulls <- data.frame()
for (i in seq(0, total_srg_rows, by = offset)) {
  if (i == 0){
    next
  }
  interim_endpoint <- paste(srg_endpnt,srg_suffix,i,sep = "")
  # print(interim_endpoint)
  interim_pull <- read.table(interim_endpoint, header = TRUE, sep = ",", dec = ".")
  running_pulls <- rbind(running_pulls,interim_pull)
}

##Putting all Srynge Pulls together
raw_syrg <- rbind(initial_pull,running_pulls)

##confirming number of total rows
# print(nrow(initial_pull))
# print(nrow(running_pulls))
# print(nrow(raw_syrg))
```

NYC Dept. City Planning Housing Database

```
## Dataset 2: NYC DCP Housing Data (Broken Down by Multiple Geographic Area's; Choosing Community District Breakdown)

### Downloading the 23Q4 CSV Zip File from the [DCP Wesbite](https://www.nyc.gov/site/planning/data-maps/open-data/dwn-housing-database.page#housingdevelopmentproject).
### Selected "HousingDB_by_CommunityDistrict.csv" from zipped contents.
download.file("https://s-media.nyc.gov/agencies/dcp/assets/files/zip/data-tools/bytes/nychousingdb_23q4_csv.zip",
              destfile = "housing_zip.zip", mode = "wb")
unzip("housing_zip.zip", files = "HousingDB_by_CommunityDistrict.csv", exdir = tempdir())
housing_path <- file.path(tempdir(), "HousingDB_by_CommunityDistrict.csv")
raw_housing <- read.table(housing_path, header = TRUE, sep = ",", dec = ".")
```

NYC Parks Disticts

```
## Dataset 3: NYC Parks DIstricts to Community District Crosswalk (https://data.cityofnewyork.us/City-Government/NYC-Parks-Districts/mebz-ditc/about_data)
crsswlk<- read.table("https://data.cityofnewyork.us/resource/mebz-ditc.csv",header = TRUE, sep = ",", dec = ".")
```

NYC Dept. Homelessness Services Individual Census by Borough, Community District, and Facility Type

```
## Pulling in Via JSON API
# Endpoint details
total_rows <- 4325 #Manually found row count on site
hmls_endpnt <- "https://data.cityofnewyork.us/resource/veav-vj3r.json"
hmls_suffix <- "?$offset="

# Initial pull to determine chunk size
initial_pull <- fromJSON(paste0(hmls_endpnt, hmls_suffix, 0)) #Fetch the first chunk

offset <- nrow(initial_pull) # Number of rows in each chunk returned by the API
if (is.null(offset) || offset == 0) {
  stop("Initial pull returned no rows. Check the API endpoint or offset logic.")
}

# List to store the chunks
running_pulls <- list()
# Loop through offsets to fetch all chunks
for (i in seq(0, total_rows, by = offset)) {
  # Construct endpoint with offset
  interim_endpoint <- paste0(hmls_endpnt, hmls_suffix, i)
  # print(paste("Fetching data from:", interim_endpoint))

  # Fetch data from the current offset
  interim_pull <- fromJSON(interim_endpoint)
  ## Ensuring Zeros instead of nulls
  interim_pull[is.na(interim_pull)] <- 0
  ## Keeping the columns we want (Only Adult Counts)
  interim_pull<-interim_pull[c("report_date","borough","community_districts","adult_shelter",
               "adult_shelter_comm_hotel", "adult_family_shelter","adult_family_comm_hotel")]
  # Check if the pull is empty (no more rows)
  if (length(interim_pull) == 0) {
    # print("No more data to fetch.")
    break
  }

  # Append the data to the list
  running_pulls[[length(running_pulls) + 1]] <- interim_pull
}

# Combine all fetched data into a single data frame
raw_hmls <- do.call(rbind, running_pulls)

# Confirming the total number of rows
 #print(paste("Total rows fetched:", nrow(raw_hmls))) ## Shuld be4325

## Filling in Borough values based on Park District
raw_hmls_processed<- raw_hmls %>%
  filter(borough!="Westchester")%>%
  mutate(cd_boro_num = case_when(borough== "Bronx"~ "2",
            borough== "Queens" ~"4",
            borough== "Manhattan" ~"1",
            borough== "Brooklyn"~"3",
            borough=="Staten Island"~"5"),
         padded_cd = str_pad(community_districts, width = 2, side = "left", pad = "0"),
         data_year = year(ymd_hms(report_date))) %>%
  mutate(full_community_district = paste0(cd_boro_num,padded_cd))


## Data Type Conversion before Sum
raw_hmls_processed[, c("adult_shelter", "adult_shelter_comm_hotel", "adult_family_shelter","adult_family_comm_hote
l")] <- lapply(raw_hmls_processed[, c("adult_shelter", "adult_shelter_comm_hotel", "adult_family_shelter","adult_fa
mily_comm_hotel")], function(x) {
  as.numeric(x)})

raw_hmls_processed$AdultHomelessCount <- rowSums(raw_hmls_processed[, c("adult_shelter", "adult_shelter_comm_hote
l",
                                                        "adult_family_shelter","adult_family_comm_hotel")])
# limiting to the columns needed
raw_hmls_limited <- raw_hmls_processed[,c("data_year","borough","full_community_district","AdultHomelessCount")]
## Grouping for a 1:1 value on year and CD, averaging th values where there are multple.
hmls_final <- raw_hmls_limited %>%
  group_by(data_year,borough, full_community_district) %>%
  summarise(
    avg_homeless_count = mean(AdultHomelessCount, na.rm=TRUE) )

hmls_final<-hmls_final %>% dplyr::rename("communitydistrict"="full_community_district",
                                  "year"="data_year")
```

Processing

Starting with Syringe Data and Geographic Crosswalk

Flattening the crosswalk longways. Multiple Community District Values in for a singular Parks District Row. Once flattened, the table can be used to for Park Districts to Community Districts to enrich the Syringe Data with Community District.

```
## Firstly, limiting to the columns i need.
crsswlk_lim <- crsswlk[,c("borough","communityboard","parkdistrict")]

## The NYC Parks DIstricts are not the same (1:1) as the Community Districts. Multiple community Districts for each
park district. Need to flatten.
crsswlk_parsed <- data.frame()
for (i in 1:nrow(crsswlk_lim)) {
  row <- crsswlk_lim[i, ]
  cd_raw_char <- as.character(row$communityboard)
  split_values <- substring(cd_raw_char, seq(1, nchar(cd_raw_char), 3), seq(3, nchar(cd_raw_char), 3))

  expanded_df <- data.frame(
    communitydistrict = split_values,
    parkdistrict = row$parkdistrict,
    brough = row$borough
  )
  crsswlk_parsed <- rbind(crsswlk_parsed, expanded_df)}

# Sorting DF
crsswlk_parsed <- crsswlk_parsed %>% arrange(communitydistrict)
```

Grouping Data to get Total Counts of Community District per Parks District

```
## Getting Denominator for Aggregate Park District Syringe Totals; The number of Community Districts within each Pa
rk District.
crsswlk_div_num_cd <- crsswlk_parsed %>%
                    group_by(parkdistrict) %>%
                    summarise(cd_count_for_pd = n_distinct(communitydistrict))

## Crosswalk Final; CD Counts added, and data grouped.
crosswalk_final <- merge(crsswlk_parsed,crsswlk_div_num_cd, by = "parkdistrict", all = FALSE)
```

Further Cleaning, Null Removal, and Limiting to what is Needed

```
## Removing those entries that have no value for a park district
# print(nrow(raw_syrg)) #35,155
syringe_lim <- raw_syrg %>% filter(district != "")

## Filling in Borough values based on Park District
syringe_lim<- syringe_lim %>%
 mutate(borough = ifelse(
    borough == "",  # Check if `category` is blank
    case_when(
      substr(district, 1, 1) == "X" ~ "Bronx",
      substr(district, 1, 1) == "M" ~ "Manhattan",
      substr(district, 1, 1) == "Q" ~ "Queens",
      substr(district, 1, 1) == "B" ~ "Brooklyn",
      substr(district, 1, 1) == "R" ~ "Staten Island",
    ),
    borough
  ))

#print(nrow(syringe_lim)) #35,060 (95 Rows Removed)

### Limiting the raw syringe data to the columns i need for my analysis
syringe_lim <- syringe_lim[,c("year","group","location","borough","district","property_type","ground_syringes","kiosk_syringes","total_syringes")]
```

Syringe Approximation by Community District

```
syringe_grouped <- syringe_lim %>%
  group_by(year,borough, district) %>%
  summarise(
    total_syringes = sum(total_syringes, na.rm=TRUE),
    ground_syringes= sum(ground_syringes, na.rm=TRUE)
  )
# print(head(syringe_grouped))

#renaming columns as needed
syringe_grouped <- syringe_grouped %>% dplyr::rename(parkdistrict = district)

## Looking at min and max years
# print(min(syringe_grouped$year))#2017
# print(max(syringe_grouped$year))#2024

syringe_grouped_enr <- merge(syringe_grouped, crosswalk_final, by = "parkdistrict", all = FALSE)
syringe_massaged <- syringe_grouped_enr %>% mutate(ttl_syring_est_interim =  total_syringes / cd_count_for_pd)

syringe_final <- syringe_massaged %>%
  group_by(year,borough,communitydistrict) %>%
  summarise(total_syringe_ests = sum(ttl_syring_est_interim, na.rm=TRUE) )
```

Starting with Housing Data

Summarizing data for total new Units by CD from the DOB for various rolling 5 years periods for each of the years within the syringe data

```
## Limiting Columns for is needed for analysis
raw_housing<-raw_housing[,c("commntydst","comp2010","comp2011","comp2012","comp2013","comp2014","comp2015","comp201
6","comp2017",
               "comp2018","comp2019","comp2020","comp2021","comp2022","comp2023")]

## Summarizing data for totals for various rolling 5 years periods for each of the years within the syringe data
###Syringe Rolling Years
housing_agg_df <- data.frame()
for (i in 1:nrow(syringe_grouped)) {
  row <- syringe_grouped[i, ]
  # row$year
  year_range <- as.character(as.integer(seq(from = (row$year-5), to = (row$year-1), length.out = 5)))
  # print(year_range)
  temp_housing <- raw_housing[,c(
                                 "commntydst",
                                 glue("comp{year_range[1]}"),
                                 glue("comp{year_range[2]}"),
                                 glue("comp{year_range[3]}"),
                                 glue("comp{year_range[4]}"),
                                 glue("comp{year_range[5]}"))]
  temp_housing$syringe_yr <-row$year
  temp_housing <- temp_housing %>% dplyr::rename("comp_yr1" = glue("comp{year_range[1]}"),
                                          "comp_yr2" = glue("comp{year_range[2]}"),
                                          "comp_yr3" = glue("comp{year_range[3]}"),
                                          "comp_yr4" = glue("comp{year_range[4]}"),
                                          "comp_yr5" = glue("comp{year_range[5]}"))
  housing_agg_df <- rbind(housing_agg_df, temp_housing)
}

#Looking at result
#print(head(housing_agg_df))

# Relative Yearly Sums for individual years within 5 yr window
housing_sums <- housing_agg_df %>%
  mutate(housingsum = comp_yr1 + comp_yr2 + comp_yr3 + comp_yr4 + comp_yr5,
    yoy_yr2 = ifelse(comp_yr1 == 0, NA, (comp_yr2 - comp_yr1) / comp_yr1 * 100),
    yoy_yr3 = ifelse(comp_yr2 == 0, NA, (comp_yr3 - comp_yr2) / comp_yr2 * 100),
    yoy_yr4 = ifelse(comp_yr3 == 0, NA, (comp_yr4 - comp_yr3) / comp_yr3 * 100),
    yoy_yr5 = ifelse(comp_yr4 == 0, NA, (comp_yr5 - comp_yr4) / comp_yr4 * 100)) %>%
  mutate(
    avg_yoy_change = rowMeans(select(., yoy_yr2, yoy_yr3, yoy_yr4, yoy_yr5), na.rm = TRUE)
  )

housing_final <-housing_sums[,c("syringe_yr","commntydst","housingsum","avg_yoy_change")] %>%
                dplyr::rename("communitydistrict"="commntydst")
```

Putting it all Together; the Final DF

```
## Joining in the homeless data to the syringe data (full join for maps. Will limit for regressions later.)
syringe_homeless_fnl<-merge(syringe_final,hmls_final, by = c('year',"borough","communitydistrict"), all = TRUE)

housing_final_join <-housing_final %>% dplyr::rename("year"="syringe_yr")

# checking data types
syringe_homeless_fnl$year <- as.integer(syringe_homeless_fnl$year)
syringe_homeless_fnl$communitydistrict <- as.integer(syringe_homeless_fnl$communitydistrict)
housing_final_join$year <- as.integer(housing_final_join$year)
housing_final_join$communitydistrict <- as.integer(housing_final_join$communitydistrict)

housing_syringe_df <- merge(syringe_homeless_fnl,
                           housing_final_join,
                           by= c("year","communitydistrict"),all=TRUE)

housing_syringe_df<-housing_syringe_df%>%distinct()

data_final <- housing_syringe_df[,c("year","borough","communitydistrict",
                                    "total_syringe_ests","avg_homeless_count",
                                    "housingsum","avg_yoy_change")]%>%
  dplyr::rename("prev5_avg_yoy_change"="avg_yoy_change")
```

Analysis

Firstly, let’s take a look at the data itself and see what the processed data shows us about these three dimensions of data.

Basic Summary Statistics

SUmmary for Syringes

```
print(summary(data_final$total_syringe_ests))
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's 
##      0.5      7.0     55.0   4875.9   1540.8 101547.0      374
```

Summary for Avg. Homeless

```
print(summary(data_final$avg_homeless_count))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##    0.00   35.33  290.92  364.35  583.70 1681.44     157
```

Summary for Housing Sum

```
print(summary(data_final$housingsum))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  -194.0   273.0   877.5  1624.7  2174.5 10451.0
```

Summary for Housing YOY Change

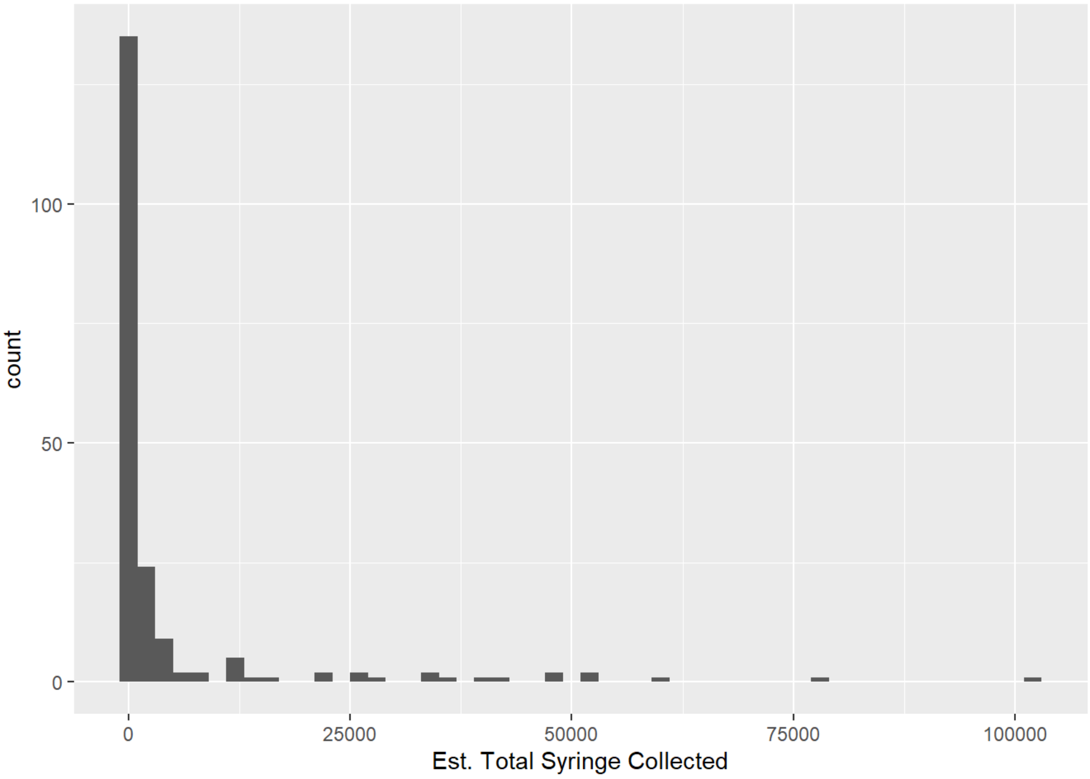
```
print(summary(data_final$prev5_avg_yoy_change))
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's 
## -2973.248     4.368    38.269   232.867   110.634 22137.460        97
```

Histograms for Distributions

Syringe Data

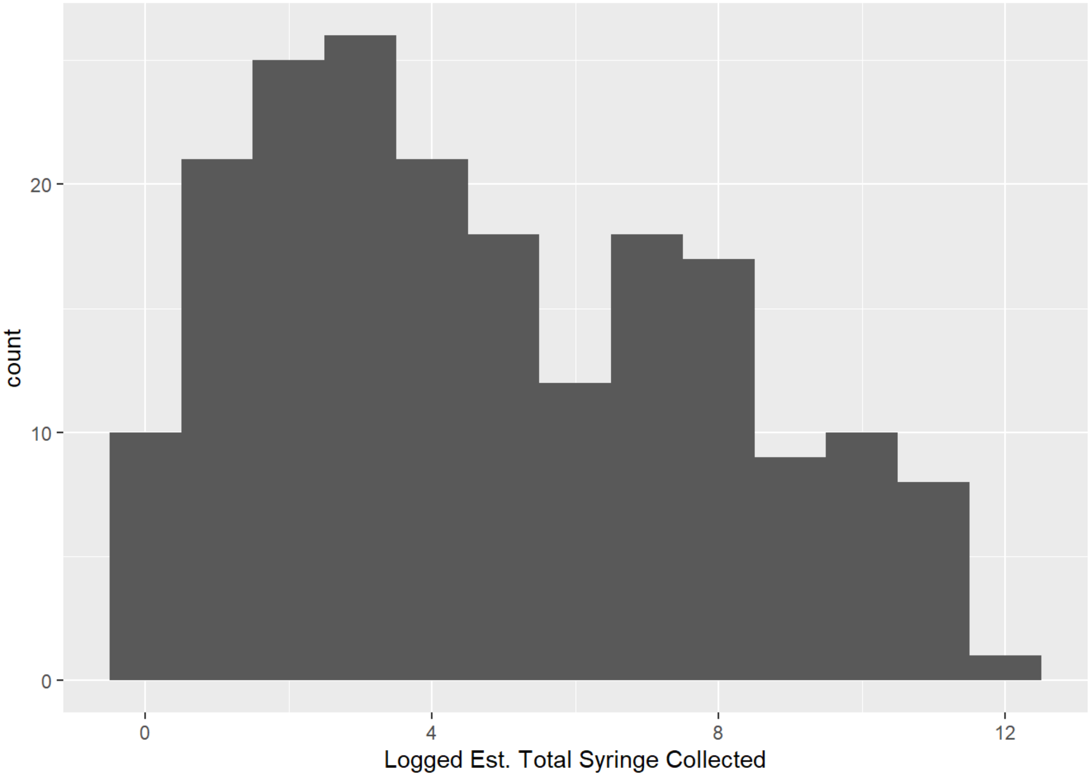
```
ggplot(data_final, aes(x = total_syringe_ests)) + geom_histogram(binwidth = 2000)+ xlab("Est. Total Syringe Collected")
```



Has a floor at Zero, cant have negative Syringes. Log Transform Needed.

Logged Syringe Data

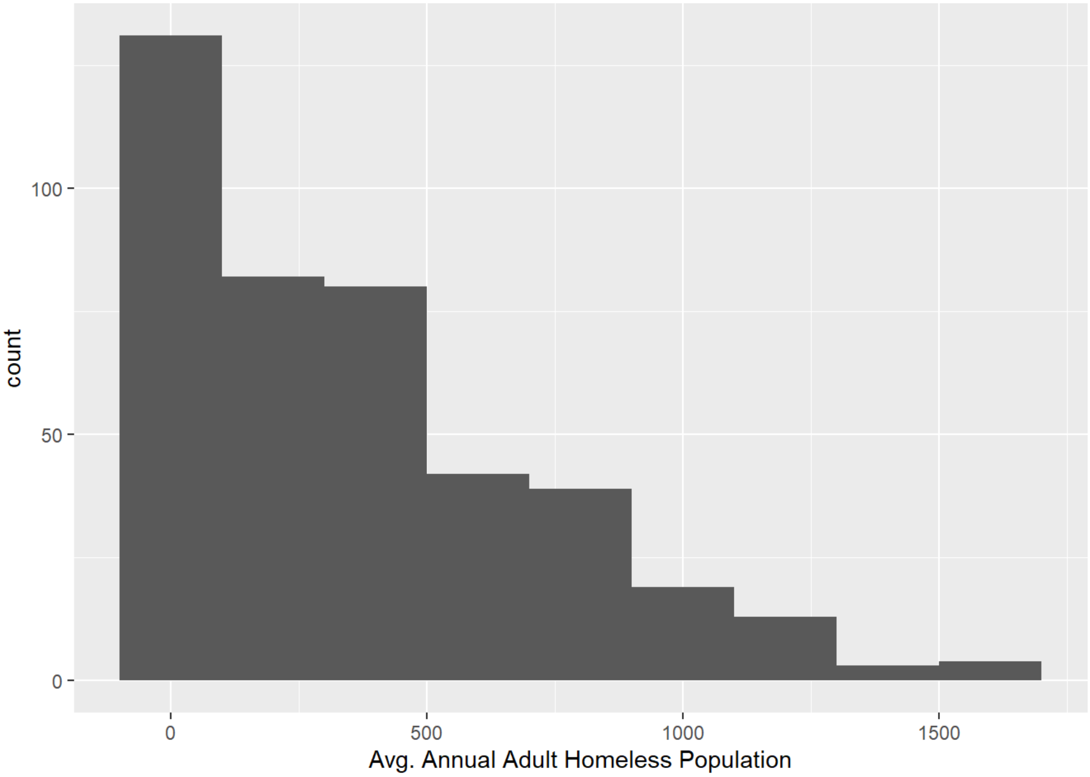
```
ggplot(data_final, aes(x = log(total_syringe_ests+1))) + geom_histogram(binwidth = 1) + xlab("Logged Est. Total Syringe Collected")
```



Looks better than original.

Adult Homeless Population

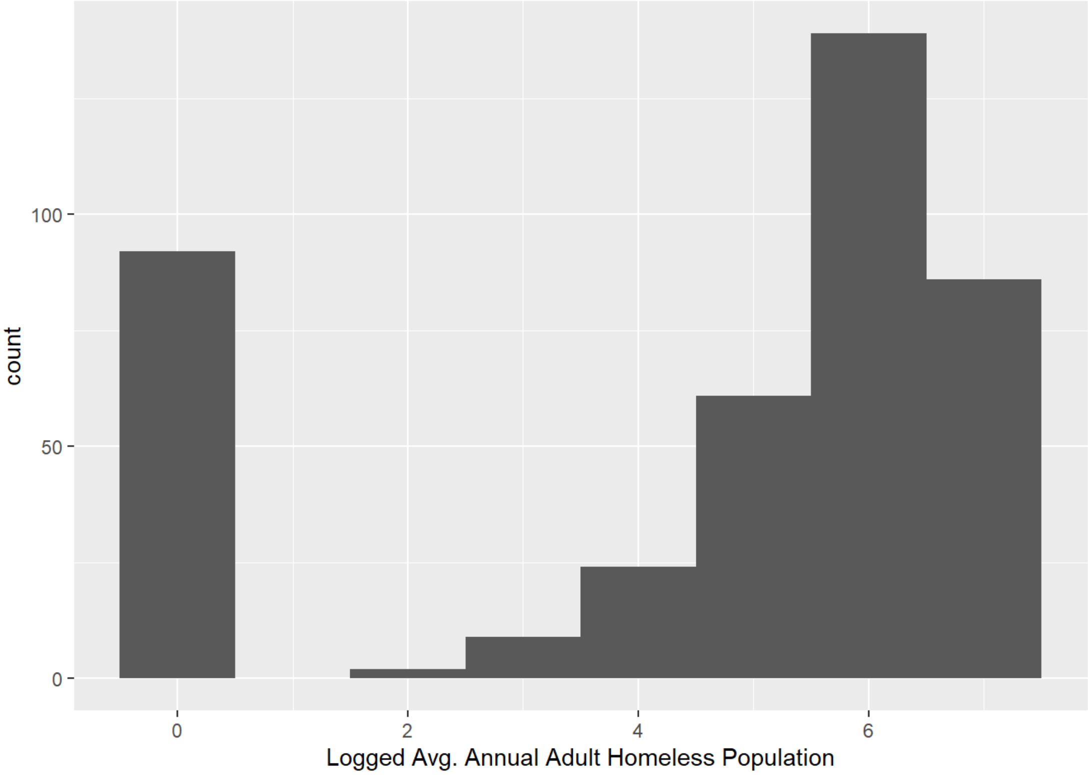
```
ggplot(data_final, aes(x = avg_homeless_count)) + geom_histogram(binwidth = 200)+ xlab("Avg. Annual Adult Homeless
Population")
```



Has a floor at zero b/c cant have negative people. Log Transform needed.

Logged Adult Homeless Population

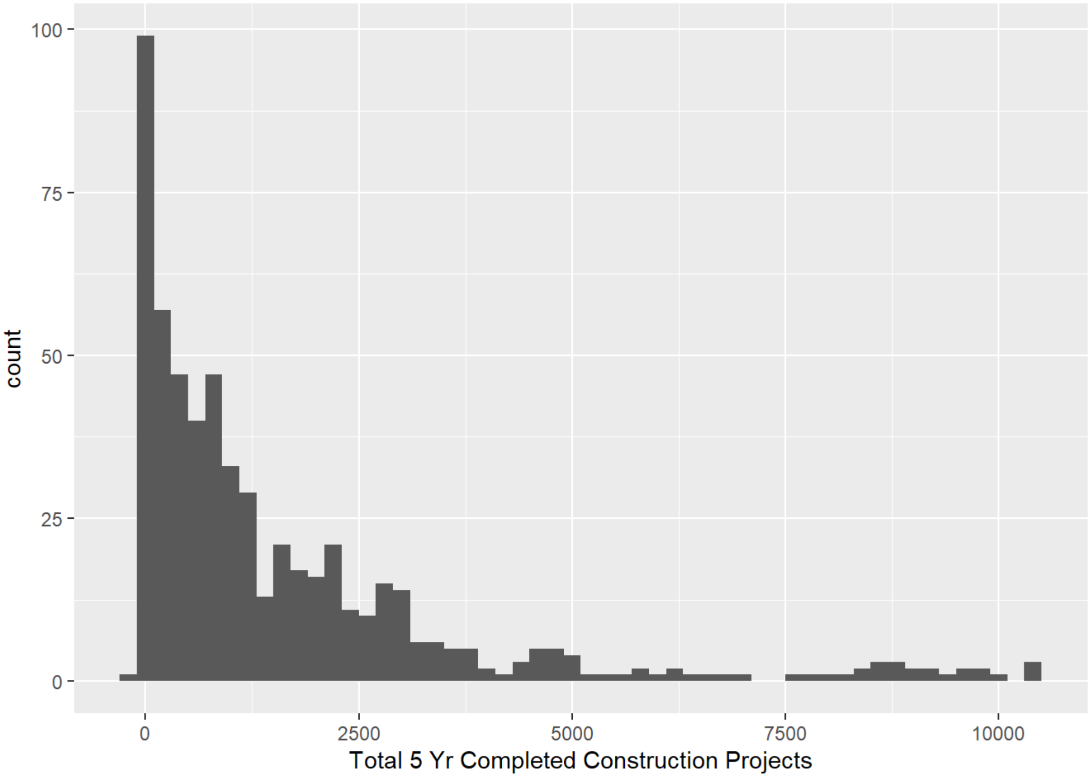
```
ggplot(data_final, aes(x = log(avg_homeless_count+1))) + geom_histogram(binwidth = 1)+xlab("Logged Avg. Annual Adult Homeless Population")
```



Looks better.

Total 5 Yr Completed Construction Projects

```
ggplot(data_final, aes(x = housingsum)) + geom_histogram(binwidth = 200) + xlab("Total 5 Yr Completed Construction Projects")
```

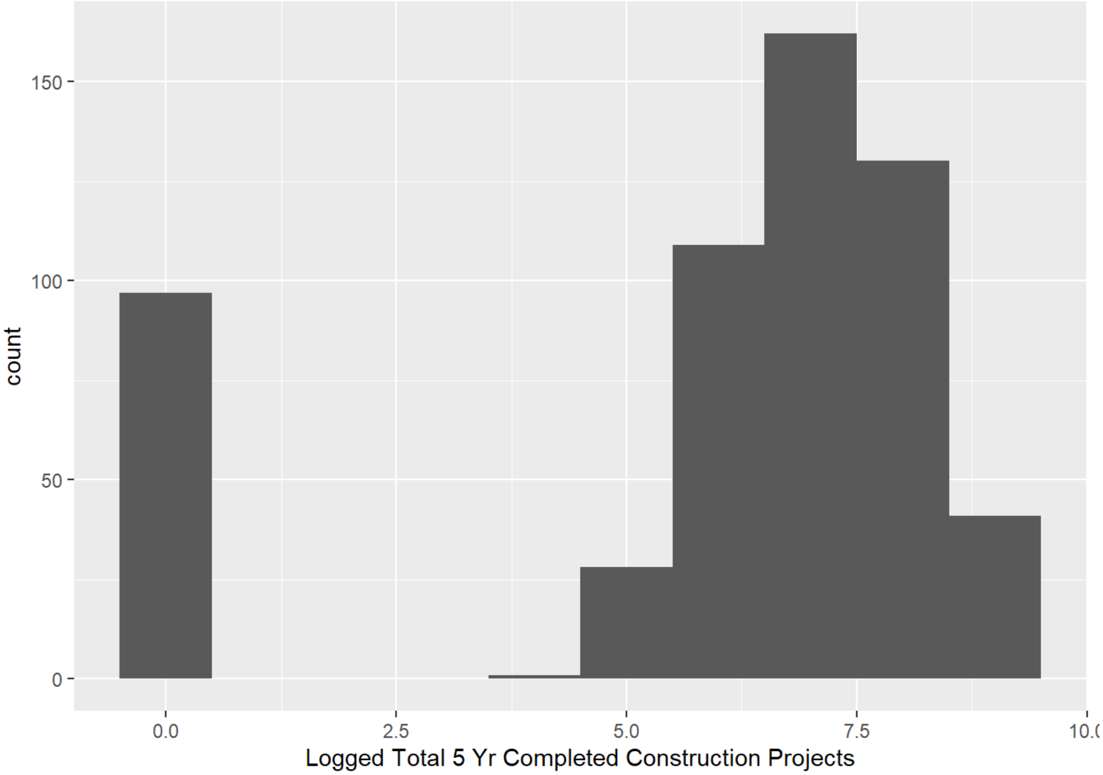


Has clustering at zero,

trying Log Transform.

Logged Total 5 Yr Completed Construction Projects

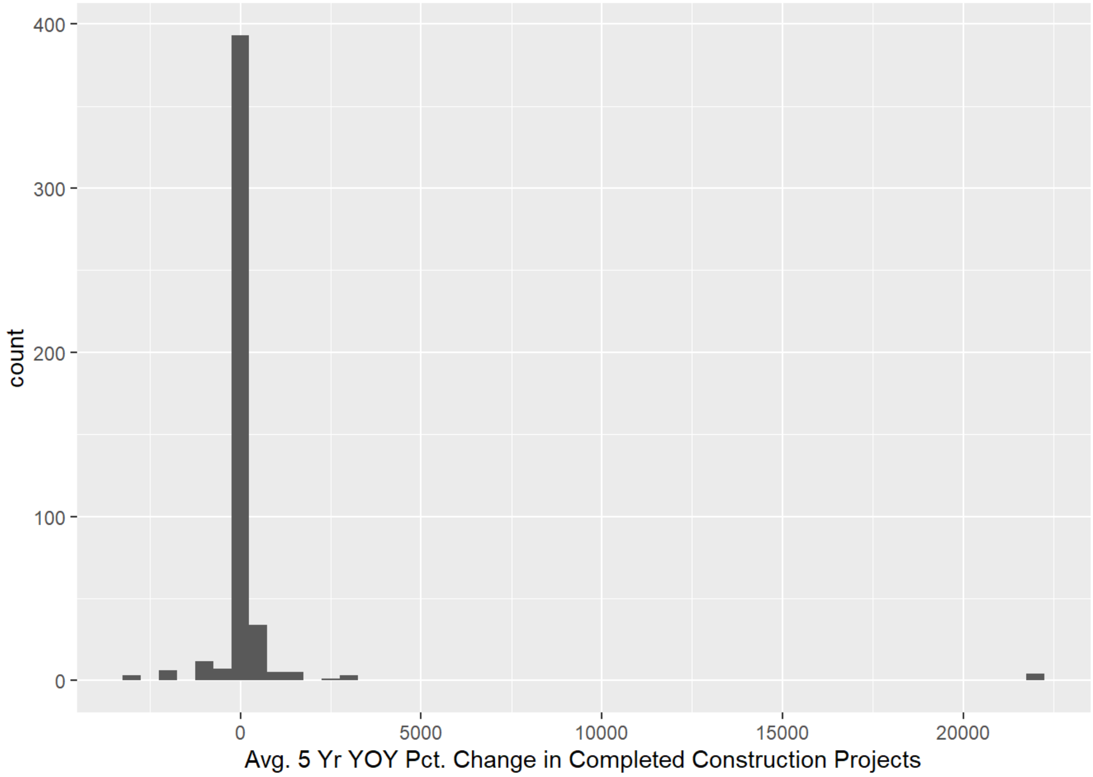
```
ggplot(data_final, aes(x = log(housingsum+1))) + geom_histogram(binwidth = 1) + xlab("Logged Total 5 Yr Completed C
onstruction Projects")
```



Still high count at 0, best I can do.

Avg. 5 Yr YOY Pct. Change in Completed Construction Projects

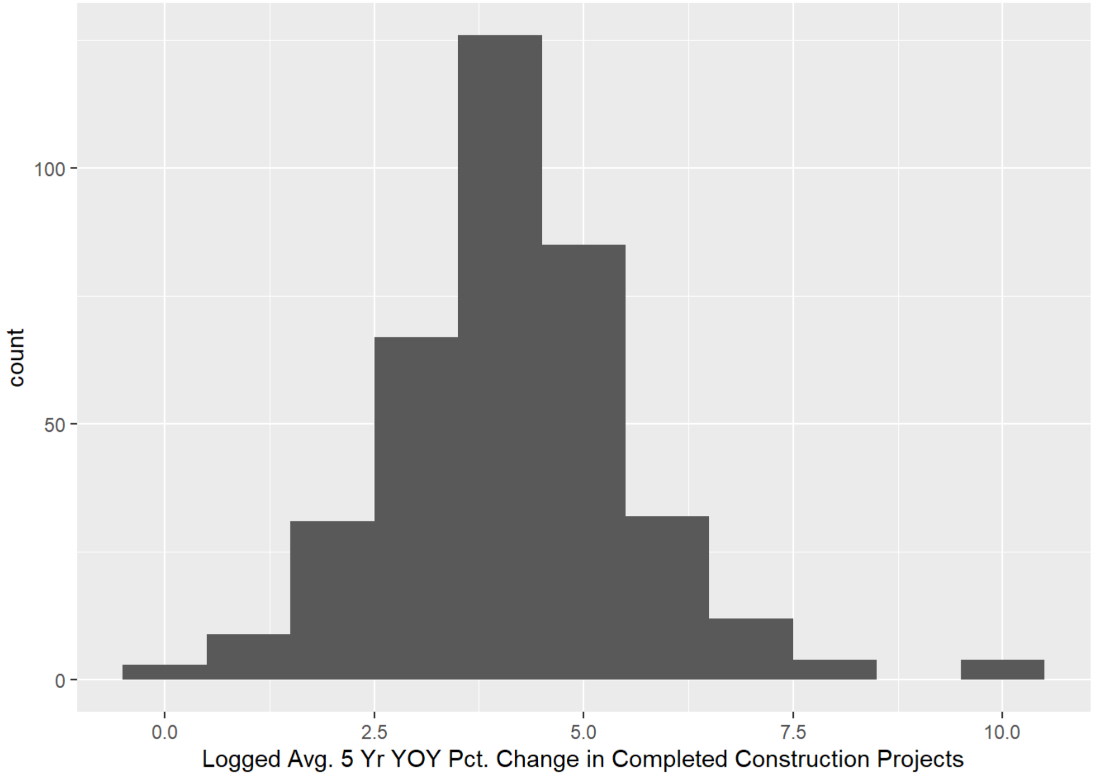
```
ggplot(data_final, aes(x = prev5_avg_yoy_change)) + geom_histogram(binwidth = 500)+xlab("Avg. 5 Yr YOY Pct. Change
in Completed Construction Projects")
```



Checking the log distribution.

Logged Avg. 5 Yr YOY Pct. Change in Completed Construction Projects

```
ggplot(data_final, aes(x = log(prev5_avg_yoy_change+1))) + geom_histogram(binwidth = 1) +xlab("Logged Avg. 5 Yr YOY
Pct. Change in Completed Construction Projects")
```



Log looks better for nor. dist.

After taking a look at the histograms for each of these variables that I want to analyze, because of the zero floors for several of them and the original distributions, Log transformations provide a more normal distribution to analyze. I will use the log transformations in regressions.

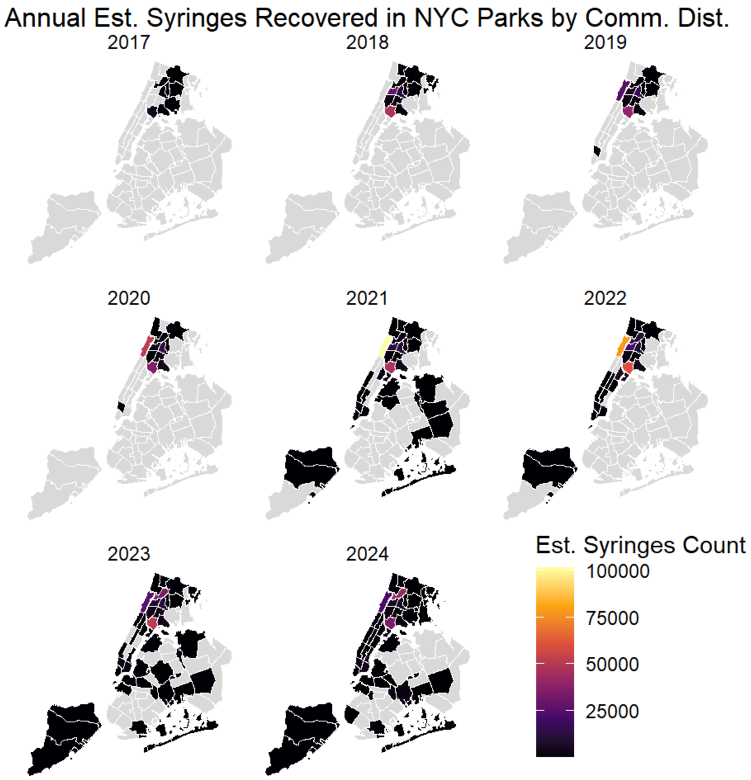
Prepping for Map Visuals

```
## Dealing with GeoJSON file in order to map data by Community District
cd_geojson <- tempfile(fileext = ".geojson")
download.file("https://data.cityofnewyork.us/api/geospatial/yfnk-k7r4?method=export&format=GeoJSON",
              cd_geojson)
## Formatting the Community DIstrict Shapefile as needed to join with the final df.
cd_sf <- read_sf(cd_geojson)
cd_sf <- cd_sf %>% dplyr::rename("communitydistrict"="boro_cd")
cd_sf$communitydistrict <- as.integer(cd_sf$communitydistrict)

### Looping through all the years in the "data_final" df to do the join for geom, to keep the blank geoms needed fo
r full map
adjusted_geom = data.frame()
for (y in unique(data_final$year)){
  lim_data <-data_final %>% filter(year==y)
  # print(length(unique(lim_data$communitydistrict)))
  # Left Joining for each year, so we have Geom for all CD, even if not in the data.
  temp_df <- cd_sf %>%
    left_join(lim_data, by = "communitydistrict")
  temp_df$year <- y
  adjusted_geom<-rbind(adjusted_geom,temp_df)
}

#Ensuring the borough Values arent null for those CD that are not in the Housing/Syringe Data
adjusted_geom <- adjusted_geom %>% mutate(borough = ifelse(is.na(borough),
                                    case_when(
                                      substr(as.character(communitydistrict), 1, 1) == "2" ~ "Bronx",
                                      substr(as.character(communitydistrict), 1, 1) == "1" ~ "Manhattan",
                                      substr(as.character(communitydistrict), 1, 1) == "4" ~ "Queens",
                                      substr(as.character(communitydistrict), 1, 1) == "3" ~ "Brooklyn",
                                      substr(as.character(communitydistrict), 1, 1) == "5" ~ "Staten Island",
                                      ),
                                    borough
                                    ))
```

Mapping To Visually See The Data in a Geospatial Context

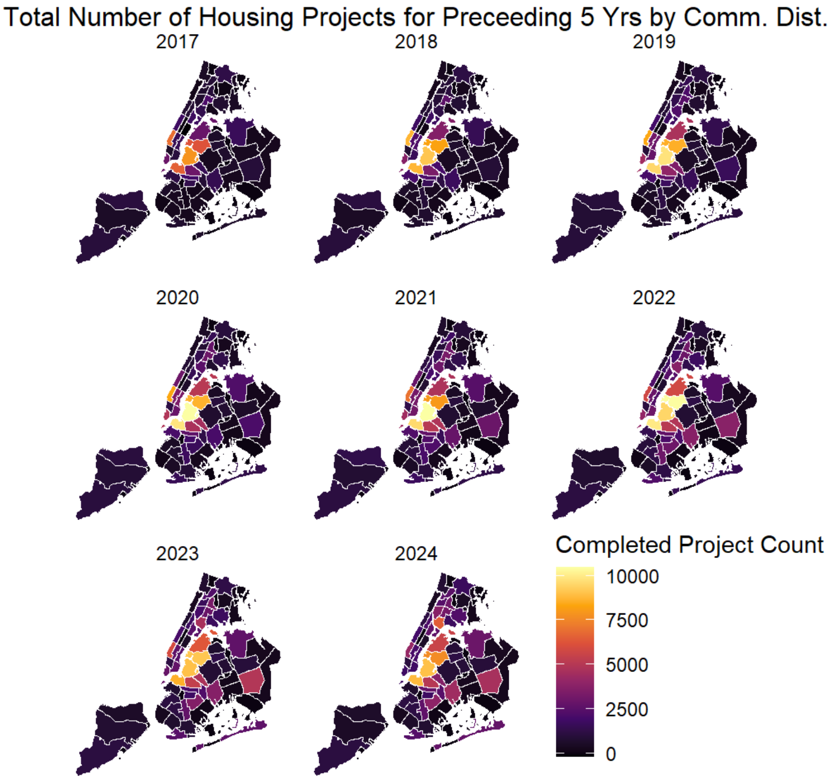
Syringe Data by Community District

Looking at the year by year break down for each syringe collection year in the data, one can identify that there are a decent number of nulls within the data. This could be from either syringes not being found by Parks staff in the vast majority of the city, or it could have to do with when and how the data started being collected. In the earlier years the syringe counts are limited to the South Bronx and small sections of Manhattan, but as the years progress additional parts of the city register a having a syringes collected.



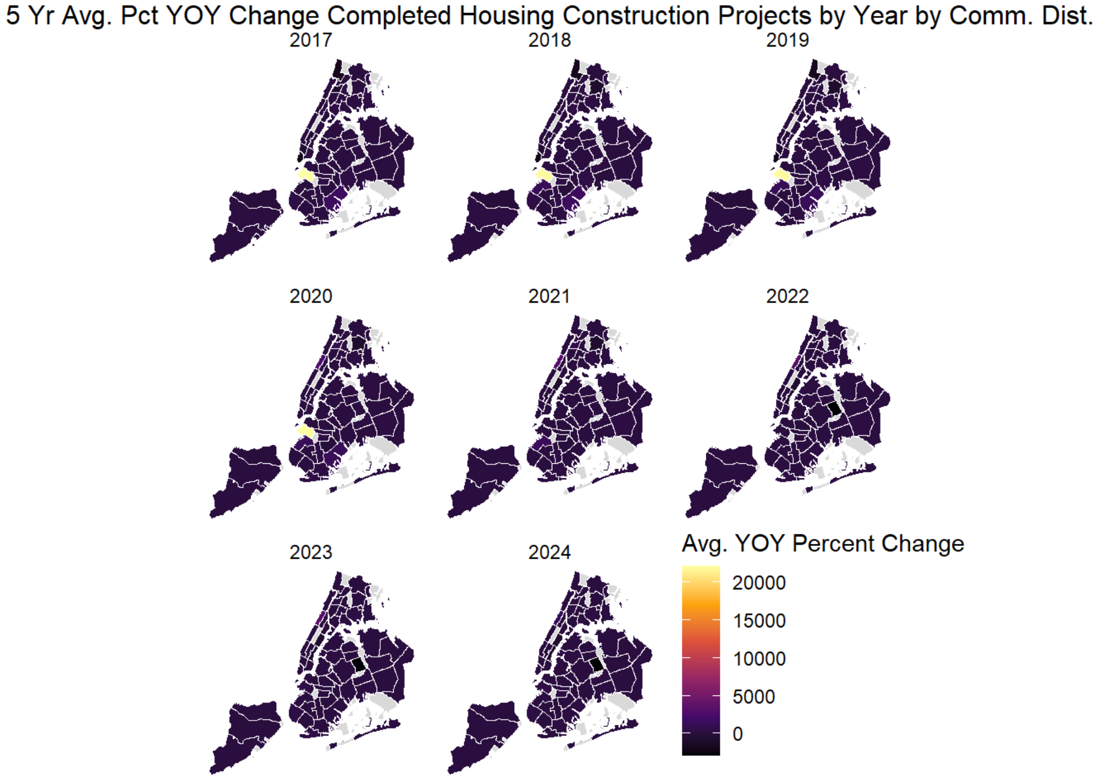
Total Number of Housing Related Construction Projects for 5 Years Previous by Community District

For the total number of housing related completed construction projects registered with DOB, in the early years the high development zones are limited to Northern Brooklyn and parts on the lower end of Manhattan. As the years progress, the number of projects through out the city, more specifically in Brooklyn and Queens, increase.



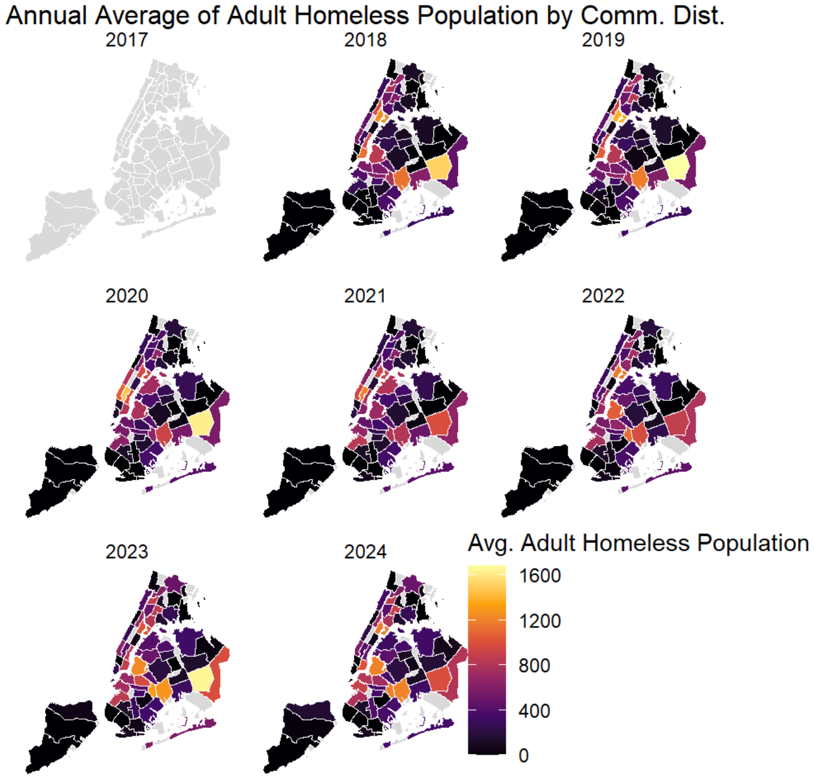
Average YOY Change for Housing Related Construction Projects 5 Years Previous by Community District

The average YOY change for completed housing related construction project is fairly consistent through out the entire city through out the years in the data. However, the areas of what looks like Sunset Park & Park Slope in Brooklyn experienced noticeably high YOY change in projects from 2017 through 2020.


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 Adult Homeless Population Based On Shelter Counts

Lastly, for the average annual adult homeless population, there is no data for 2017, but once data is collected for 2018 there are many different parts of the city that have fairly high average adult homeless populations. There doesn’t seem to be any particular trend visible just by looking at the data. However, it should be noted, that while this data is useful for gauging the homeless population in the city, at a Community District geographic boundary level, the means by which the data is collected - taking counts at homeless shelters of different kinds - implicitly causes the homeless population to be higher in areas that have hotels or shelters designated for that population. This may not necessarily be an issue, but it is something to keep in mind when comparing this variable to others.


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

Removing Nulls from Finalized DF

```
#Removing the Nulls from the final data df in preparation for regression
data_final_nonull <- na.omit(data_final)
```

Linear Regressions w/ Community District Geographies

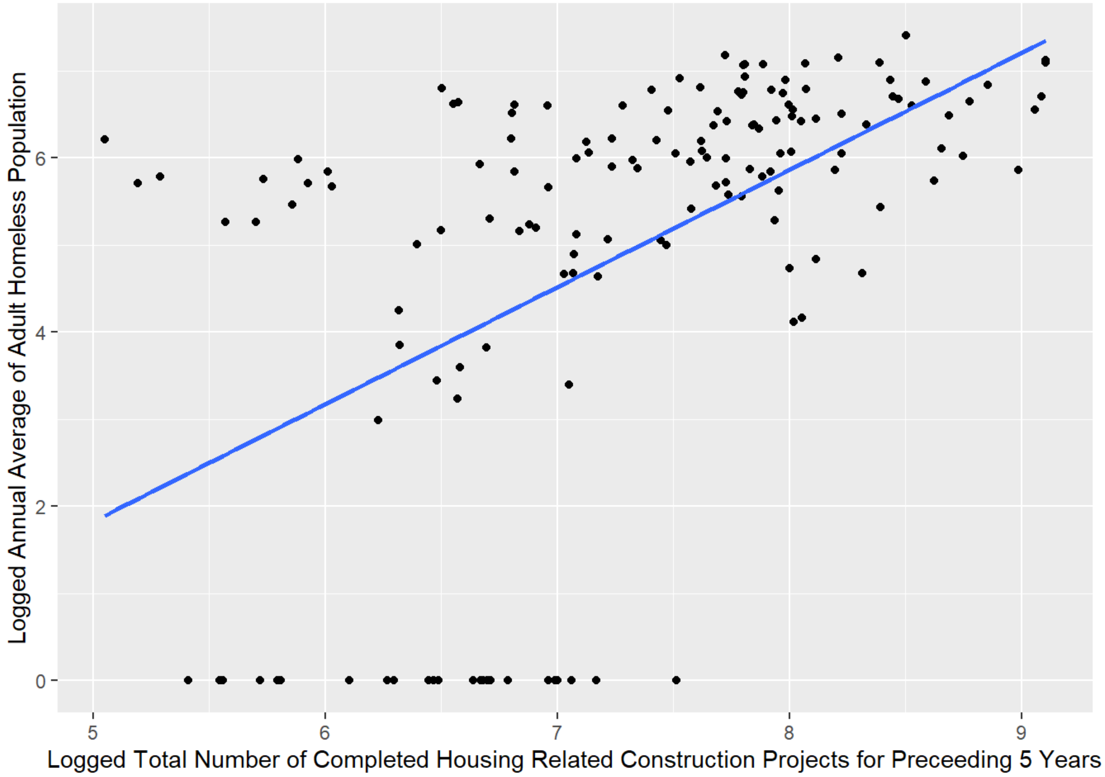
As outlined in the introduction, there are two parts to this analysis. The first is to see if there is a correlation between housing-related construction projects and the homeless population. The second is to see if the size of homeless population correlates with the amount of syringes found in NYC parks. This first attempt at regression models will maintain a larger number of data points constrained by community district. However, the wide array of factors that influence various community districts may influence the regression results obfuscating any substantial correlation. Factors such as wealthier, gentrifying areas of the city having more construction projects associated with their geographic boundary, while intravenous drug use is more likely to be concentrated in poorer, less wealthy community districts. Additionally, housing units through out the city may not properly be shown to reduce homelessness due to the homeless population data being sourced from shelters. Shelters are not evenly distributed throughout the city, making the numbers at the community district level skewed for the overall numbers.

QUESTION 1: Does the Number of Housing Related Construction Projects for the Previous 5 Years influence the Size of Adult Homeless Population?

To answer this question, a linear regression analysis will look at the relationship between the annual average for adult homeless shelter populations and the number of completed housing construction jobs on record with the DOB. For the housing numbers, we will look at the average annual year-over-year (YOY) change in housing construction projects for the five year window previous to the homeless population year, as well as the total number of cumulative housing construction projects completed in for the five years previous to the homeless population year. Lastly, as outlined previously because of the original distributions of he variable a log transform will be used in order to get a more normal distribution for the analyzed variables.

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

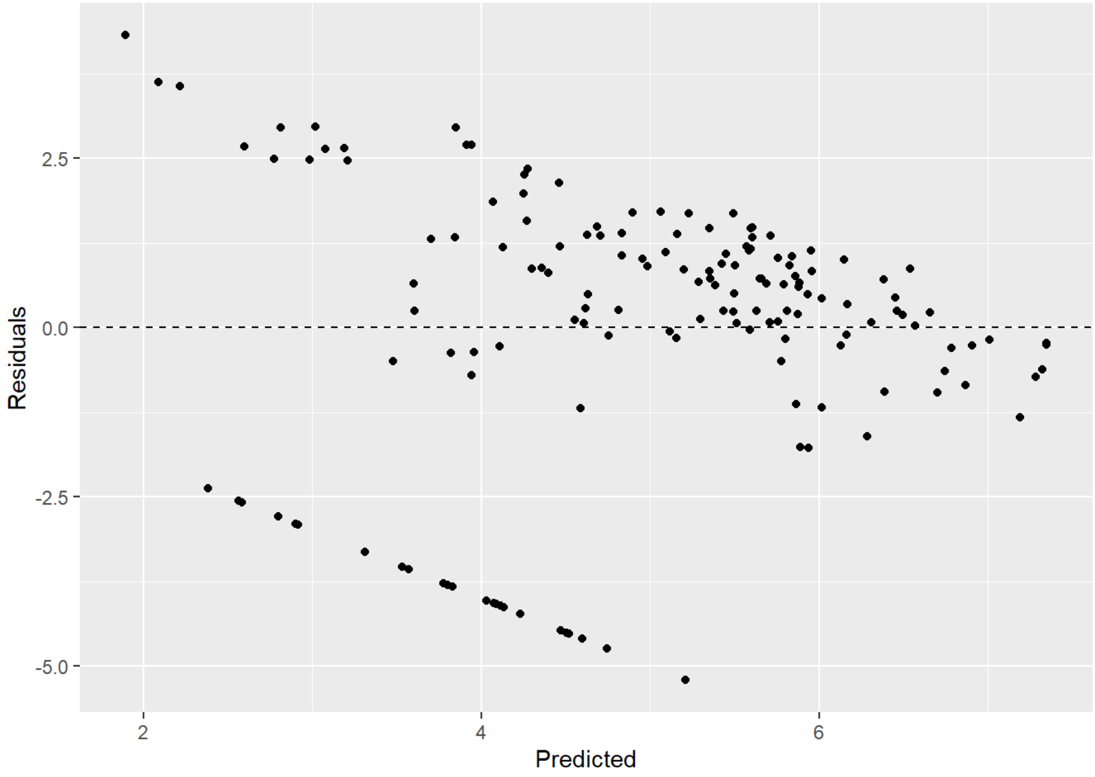
```
##
## Call:
## lm(formula = log(avg_homeless_count + 1) ~ log(housingsum + 1),
##     data = data_final_nonull)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.2104 -0.6184  0.4339  1.1790  4.3159
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (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
```



Model 1 **is** Statisitcally Significant, so Checking Model Validity.

Model 1 Linearity Check

```
# Linearity Check
ggplot(m1, aes(x=.fitted, y=.resid)) +
  geom_point()+
  geom_hline(yintercept = 0, linetype = "dashed") +
  labs(x = "Predicted",y ="Residuals")
```

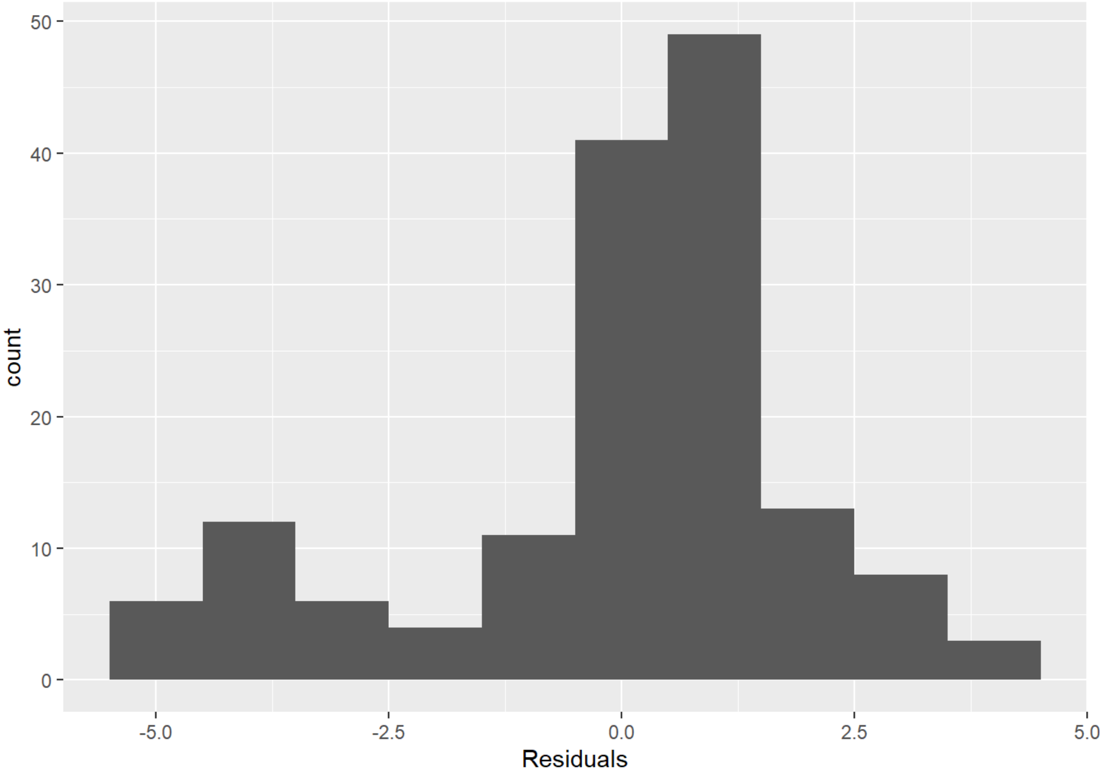


There seems to be slight

correlation in lower portion. Line seems to have Zero Slope.

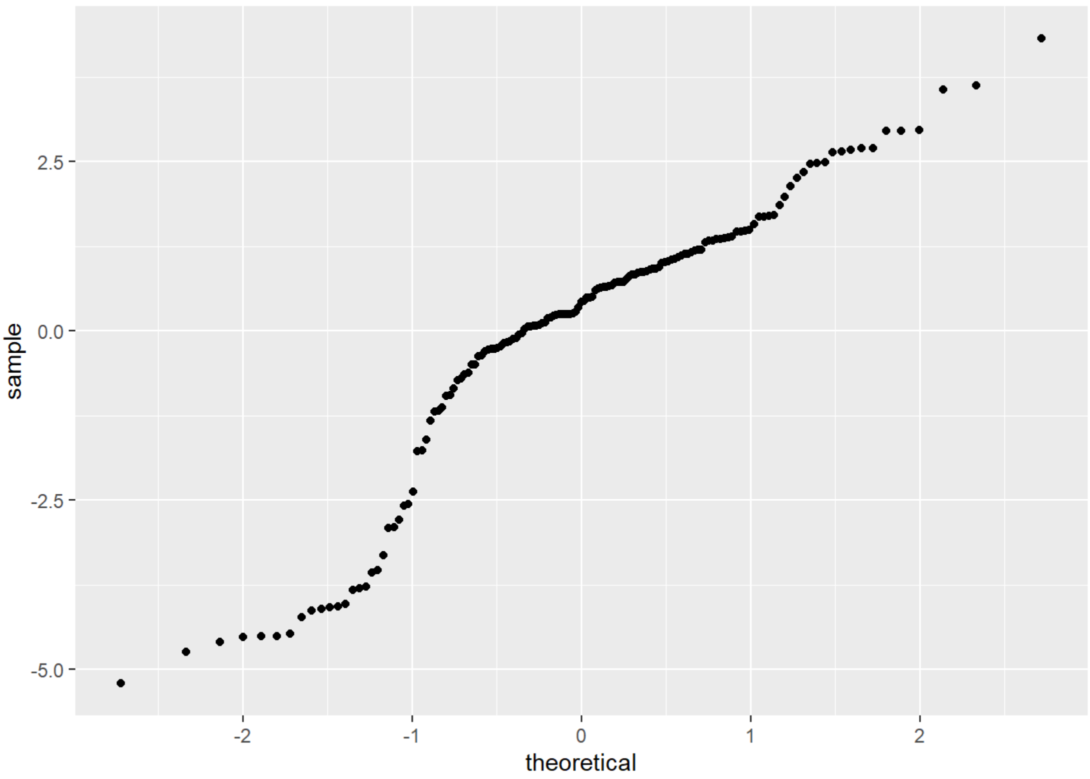
Model 1 Residual Distribution Check 1

```
# Residual Dist.
ggplot(data = m1, aes(x = .resid)) +geom_histogram(binwidth = 1) + xlab("Residuals")
```



Model 1 Residual Distribution Check 2

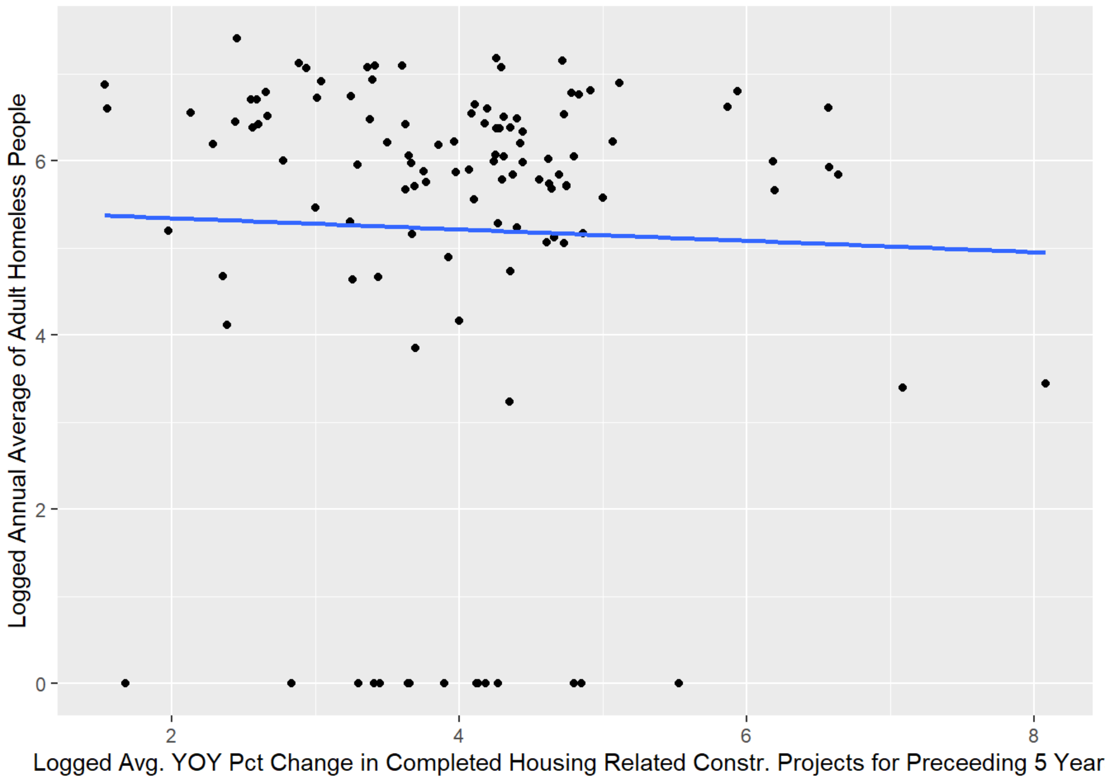
```
#Variability of Constant
ggplot(data = m1, aes(sample = .resid)) +stat_qq()
```



The residuals are basically normally distributed. Model most likely valid.

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
```

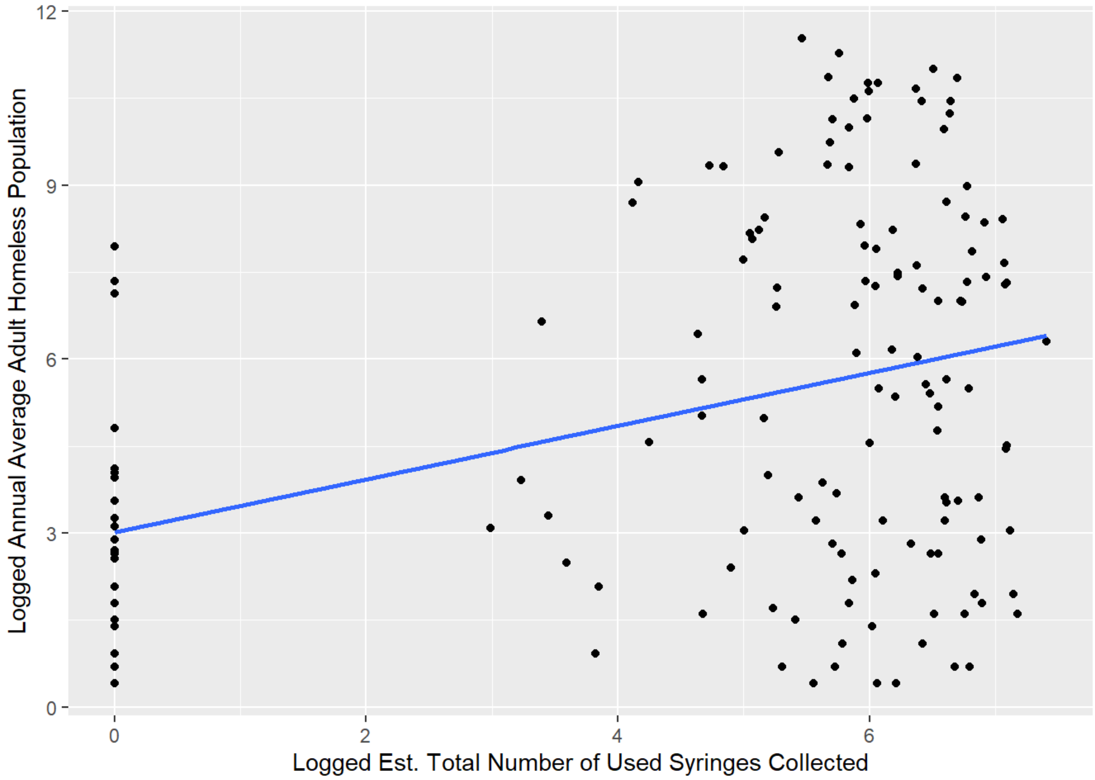


Model 2 is **NOT** Statistically Significant. Moving On.

QUESTION 2: Does the size of the Adult Homeless Population influence the Number of Used Syringes recovered by NYC Parks?

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

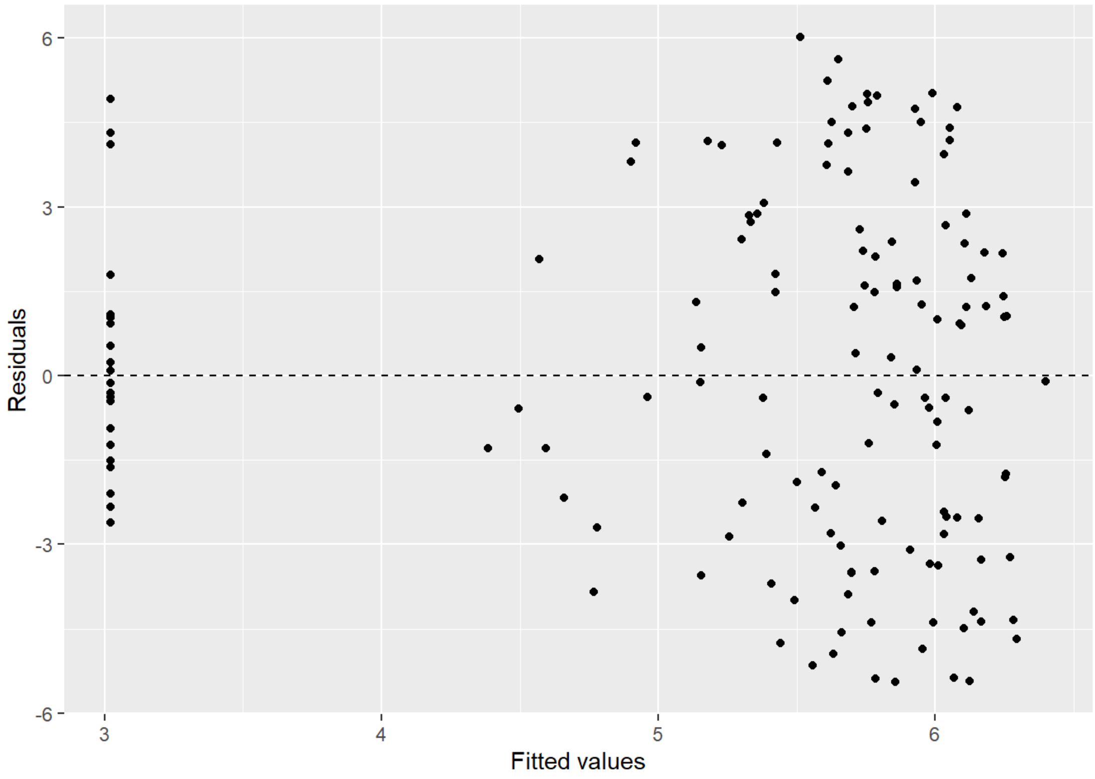
```
##
## Call:
## lm(formula = log(total_syringe_ests + 1) ~ log(avg_homeless_count +
##     1), data = data_final_nonull)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.4495 -2.5458 -0.1314  2.3448  6.0140
##
## Coefficients:
##                            Estimate Std. Error t value 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
```



Model 3 is Statisitcally Significant. Checking Validity.

Model 3 Linearity Check

```
# Linearity Check
ggplot(m3, aes(x=.fitted, y=.resid)) +
  geom_point()+
  geom_hline(yintercept = 0, linetype = "dashed") +
  labs(x = "Fitted values",y ="Residuals")
```

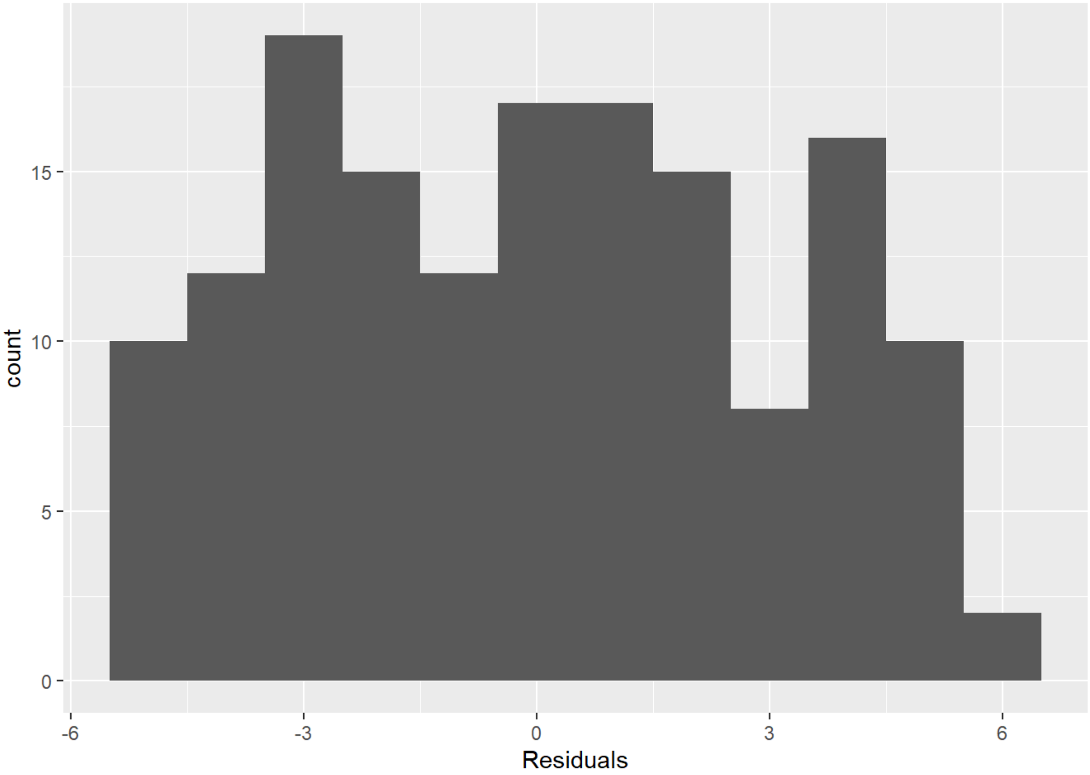


Linearity Check should be

good.

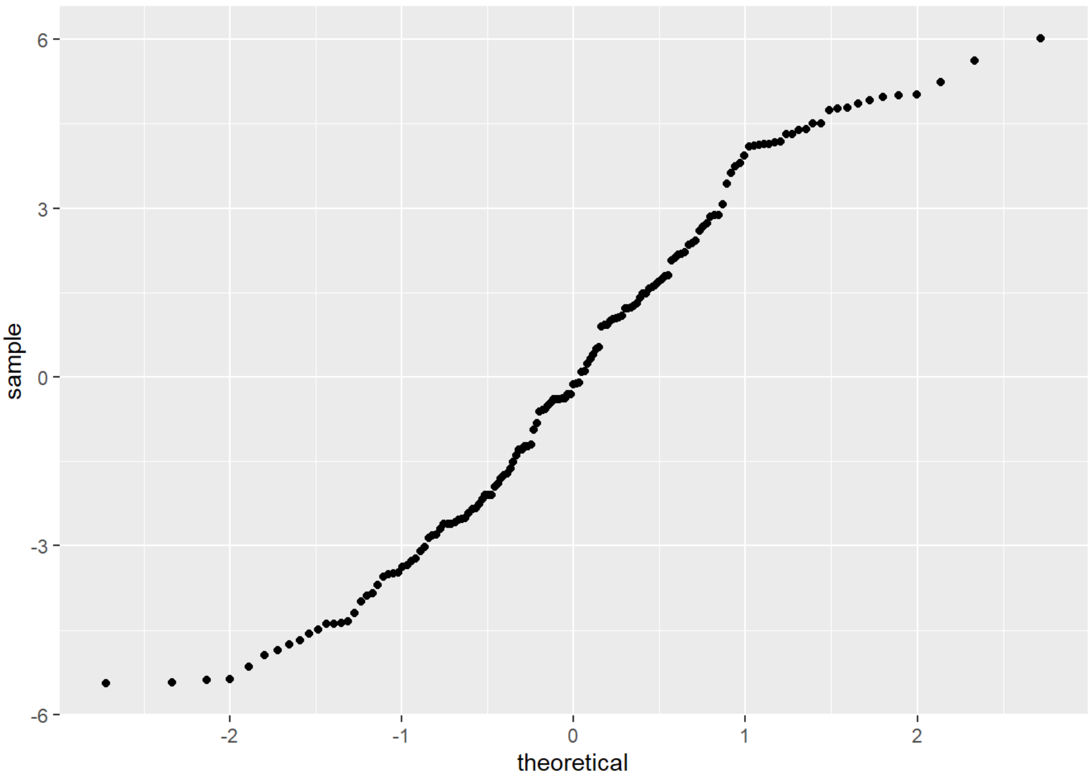
Model 3 Residual Distribution Check 1

```
# Residual Dist.
ggplot(data = m3, aes(x = .resid)) +geom_histogram(binwidth = 1) + xlab("Residuals")
```



Model 3 Residual Distribution Check 2

```
#Variablitiy of Constant
ggplot(data = m3, aes(sample = .resid)) +stat_qq()
```



```
## Slight Deviation b/c of right skew, but passes for normal.
```

Residuals are normally distributed.Model 3 most likely Valid.

Linear Regressions with Aggregate Borough Geographies

After performing the regression analysis above, the results were unremarkable. There were two statistically significant relationships. Due to potential issues with the data at the Community District level, such as an unequal distribution of parks where syringes can be found, specific types of zoning limiting where shelters exist implicitly skewing homeless populations, or housing projects being limited to select areas. I decided to aggregate to the borough level instead. Generalizing the geographic limitations may show relationships more clearly than the more granular boundaries. The main draw back to this aggregation is a reduction in data points, however it is worth taking a look.

Aggregating the data to the borough level.

Syringe Borough Aggregation

```
syringe_grouped_boro <- syringe_lim %>%
  group_by(year,borough) %>%
  summarise(
    total_syringes = sum(total_syringes, na.rm=TRUE)
    )
```

Homeless Borough Aggregation

```
raw_hmls_limited_boro <- raw_hmls_processed[,c("report_date","data_year","borough","AdultHomelessCount")]

## Grouping by report date to sum up boro total for report date.
hmls_grpd_boro <- raw_hmls_limited_boro %>%
  group_by(report_date,data_year,borough) %>%
  summarise(
    boro_sum_homeless_count = sum(AdultHomelessCount, na.rm=TRUE))

## Averaging for each boto for each year
hmls_final_boro <- hmls_grpd_boro %>%
  group_by(data_year,borough) %>%
  summarise(
    avg_homeless_count = mean(boro_sum_homeless_count, na.rm=TRUE))

hmls_final_boro <- hmls_final_boro %>% dplyr::rename(year=data_year)
```

Housing Borough Agg

```
housing_agg_df_boro <- housing_agg_df %>%
  mutate(borough = case_when(
    substr(as.character(commntydst), 1, 1) == "2" ~ "Bronx",
    substr(as.character(commntydst), 1, 1) == "4" ~ "Queens",
    substr(as.character(commntydst), 1, 1) == "3" ~ "Booklyn",
    substr(as.character(commntydst), 1, 1) == "5" ~ "Staten Island",
    substr(as.character(commntydst), 1, 1) == "1" ~ "Manhattan"))

housing_agg_df_boro_grpd<- housing_agg_df_boro %>%
  group_by(syringe_yr, borough) %>%
  summarize(
    comp_yr1 = sum(comp_yr1, na.rm=TRUE),
    comp_yr2 = sum(comp_yr2, na.rm=TRUE),
    comp_yr3 = sum(comp_yr3, na.rm=TRUE),
    comp_yr4 = sum(comp_yr4, na.rm=TRUE),
    comp_yr5 = sum(comp_yr5, na.rm=TRUE)
  )

housing_sums_boro <-   housing_agg_df_boro_grpd %>%
  mutate(housingsum = comp_yr1 + comp_yr2 + comp_yr3 + comp_yr4 + comp_yr5,
    yoy_yr2 = ifelse(comp_yr1 == 0, NA, (comp_yr2 - comp_yr1) / comp_yr1 * 100),
    yoy_yr3 = ifelse(comp_yr2 == 0, NA, (comp_yr3 - comp_yr2) / comp_yr2 * 100),
    yoy_yr4 = ifelse(comp_yr3 == 0, NA, (comp_yr4 - comp_yr3) / comp_yr3 * 100),
    yoy_yr5 = ifelse(comp_yr4 == 0, NA, (comp_yr5 - comp_yr4) / comp_yr4 * 100))

housing_sums_boro$avg_yoy_change <- rowMeans(housing_sums_boro[, c('yoy_yr2', 'yoy_yr3', 'yoy_yr4', 'yoy_yr5')])

housing_final_boro <- housing_sums_boro[,c("syringe_yr","borough","housingsum","avg_yoy_change")] %>%
                dplyr::rename("year"="syringe_yr")
```

Joining Aggregated Data Sets Together for Borough Level Geography

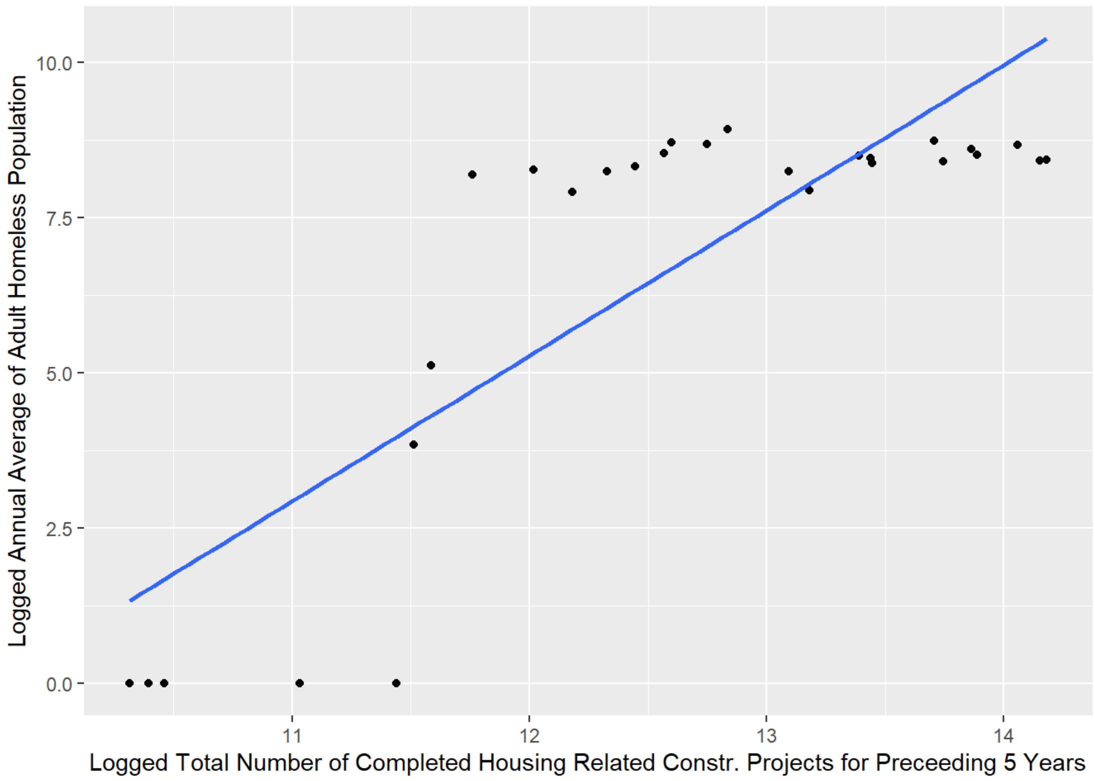
```
# first join
 syringe_homeless_boro<- merge(syringe_grouped_boro,hmls_final_boro, by = c('year',"borough"), all = TRUE)

# checking data types
syringe_homeless_boro$year <- as.integer(syringe_homeless_boro$year)
housing_final_boro$year <- as.integer(housing_final_boro$year)

#second join
housing_syringe_df_boro <- merge(syringe_homeless_boro,
                           housing_final_boro,
                           by= c("year","borough"),all=TRUE)

#To keep more data points with the Agg, making the NA syrings values zero
housing_syringe_df_boro <- housing_syringe_df_boro %>%
    mutate(total_syringes = ifelse(is.na(total_syringes), 0, total_syringes))

# Omitting the remaining nulls
data_final_boro_nonulls<- na.omit(housing_syringe_df_boro)
```

Not checking the distributions of raw variables again, as I did before. We know some have floors around zero, so I will just log the variables again.

QUESTION 1: Does the Number of Housing Related Construction Projects for the Previous 5 Years influence the Size of Adult Homeless Population?

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

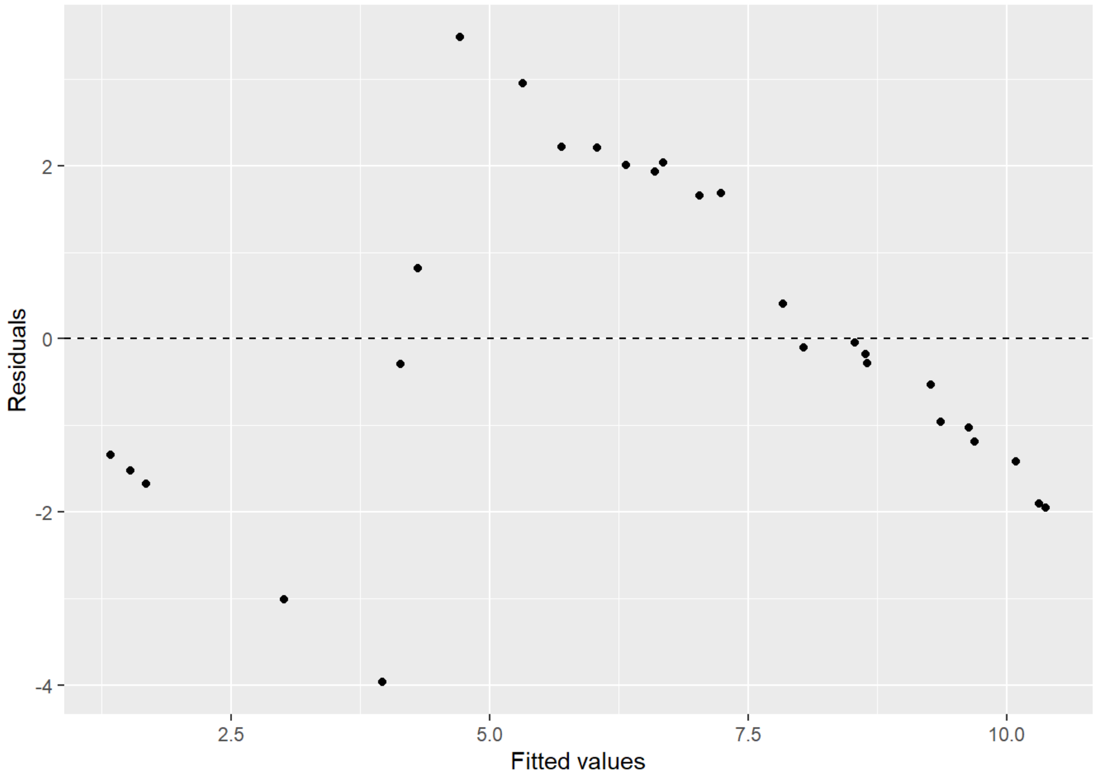
```
##
## Call:
## lm(formula = log(avg_homeless_count + 1) ~ log(housingsum + 1),
##     data = data_final_boro_nonulls)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9643 -1.3569 -0.2316  1.7439  3.4821
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (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
```



Model 4 **is** Statisitcally Significant, checking validity.

Model 4 Linearity Check

```
# Linearity Check
ggplot(m4, aes(x=.fitted, y=.resid)) +
  geom_point()+
  geom_hline(yintercept = 0, linetype = "dashed") +
  labs(x = "Fitted values",y ="Residuals")
```

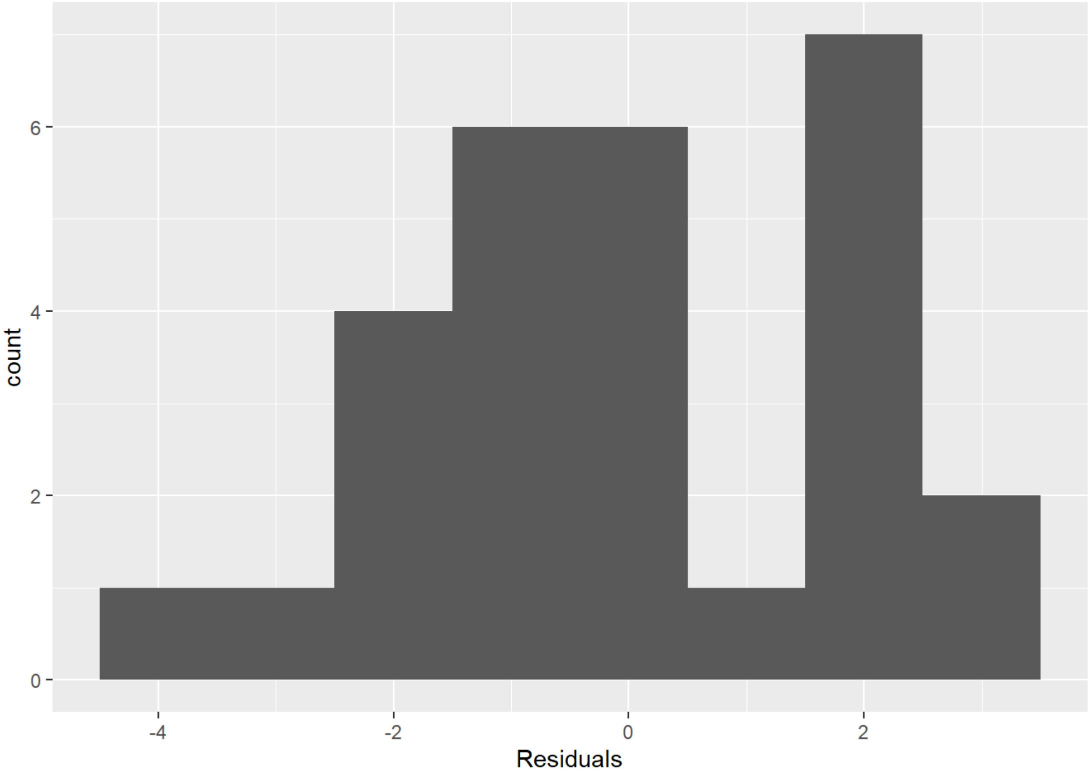


Linearity Check doesnt

seem to pass. There is a pattern in the portions of the data.

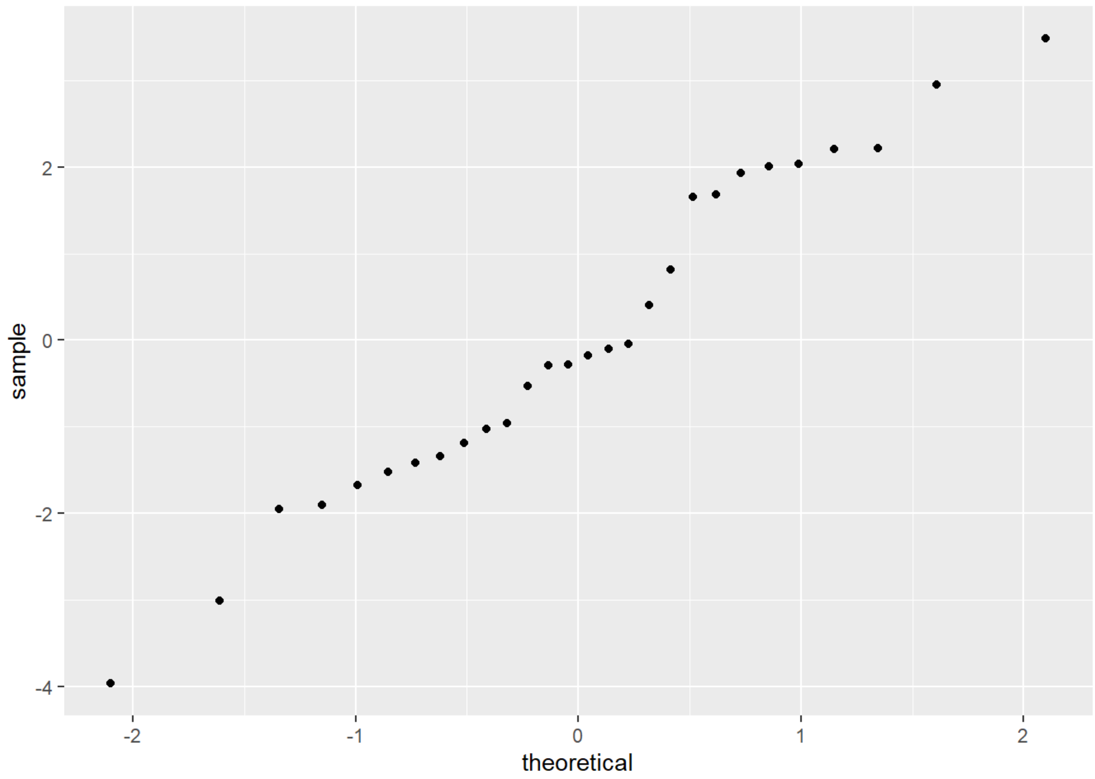
Model 4 Residual Distribution Check 1

```
# Residual Dist.
ggplot(data = m4, aes(x = .resid)) +geom_histogram(binwidth = 1) + xlab("Residuals")
```



Model 4 Residual Distribution Check 2

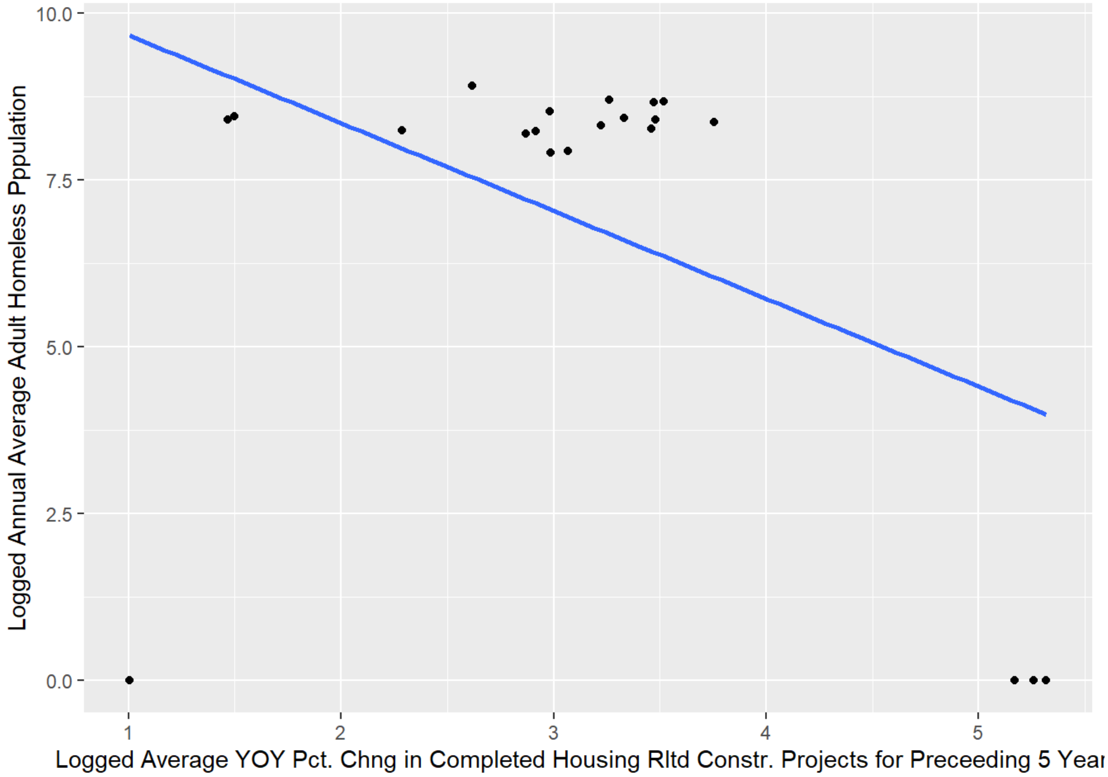
```
#Variablity of Constant
ggplot(data = m4, aes(sample = .resid)) +stat_qq()
```



MODEL 4 may not be valid. Redsiduals and Predicted values have a pattern in the points. The residuals have mostly normal distribution. Slight deviation, but normal.

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
```

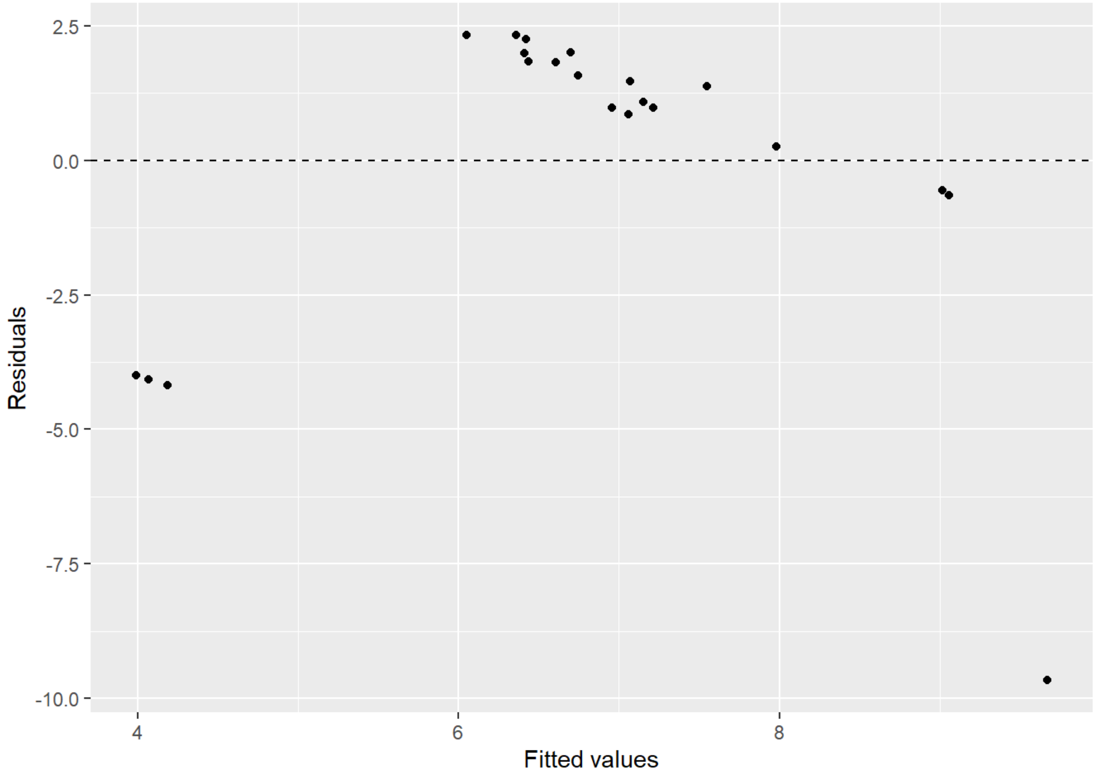


Model 5 **is** Statistic

Significant, checking validity.

Model 5 Linearity Check

```
# Linearity Check
ggplot(m5, aes(x=.fitted, y=.resid)) +
  geom_point()+
  geom_hline(yintercept = 0, linetype = "dashed") +
  labs(x = "Fitted values",y ="Residuals")
```

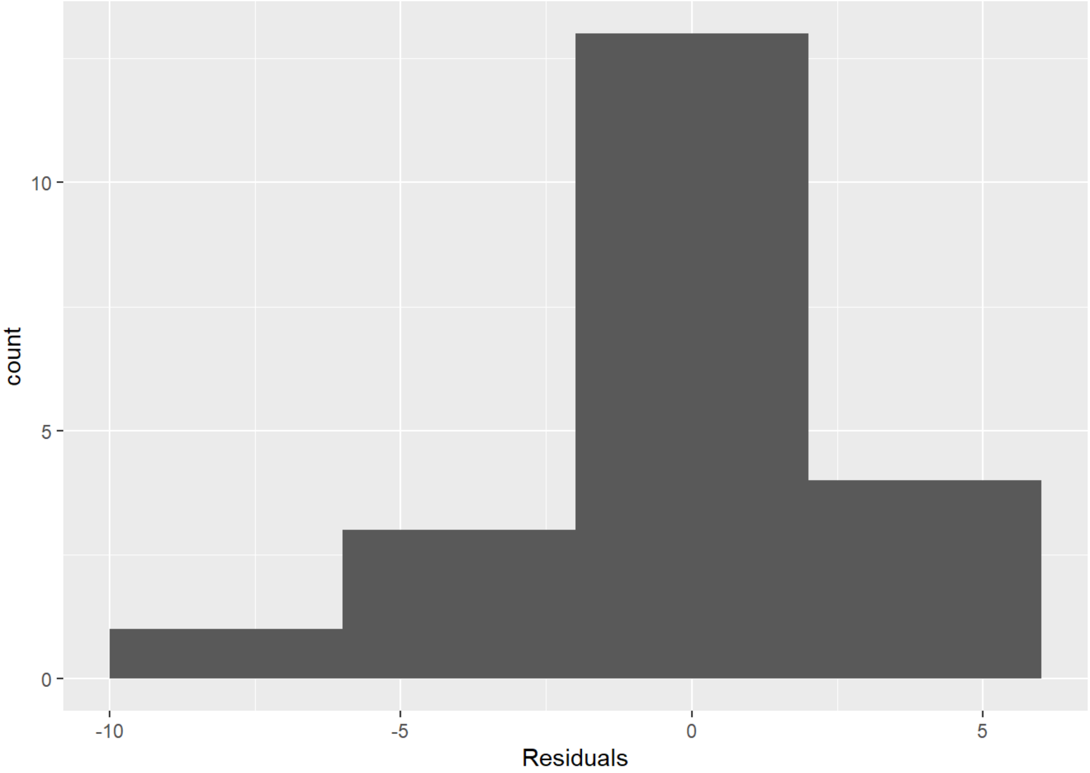


Slight pattern with outliers.

May skew regression.

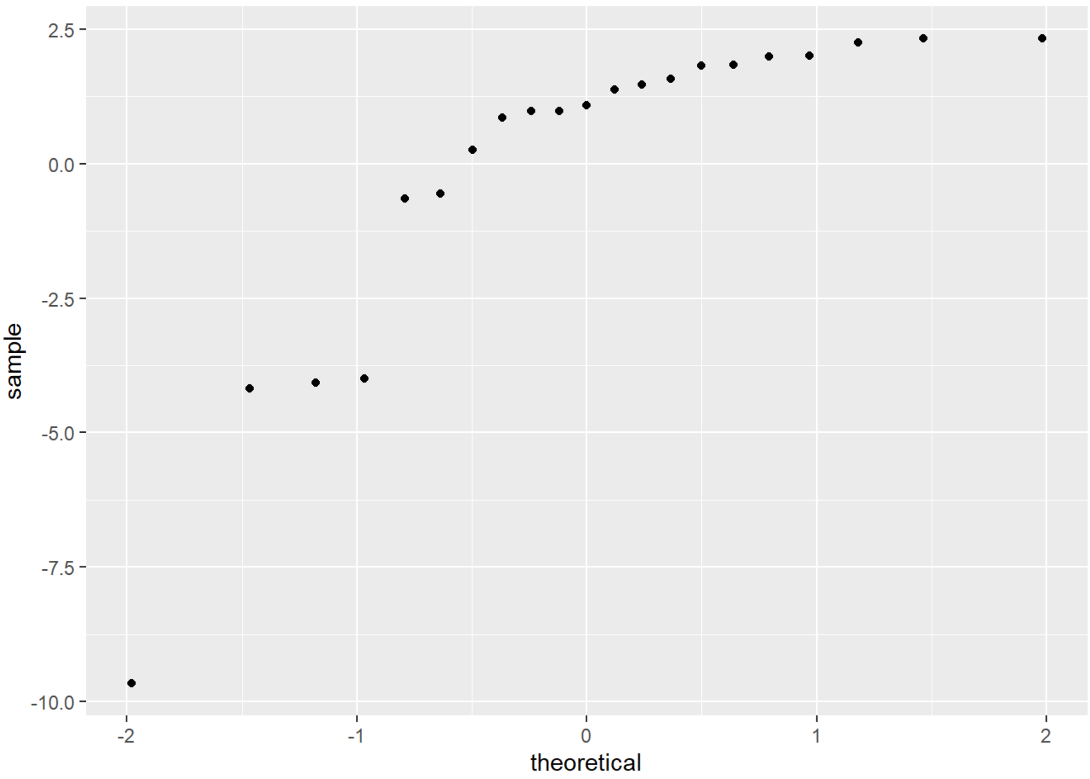
Model 5 Residual Distribution Check 1

```
# Residual Dist.
ggplot(data = m5, aes(x = .resid)) +geom_histogram(binwidth = 4) + xlab("Residuals")
```



Model 5 Residual Distribution Check 2

```
#Variablitiy of Constant
ggplot(data = m5, aes(sample = .resid)) +stat_qq()
```

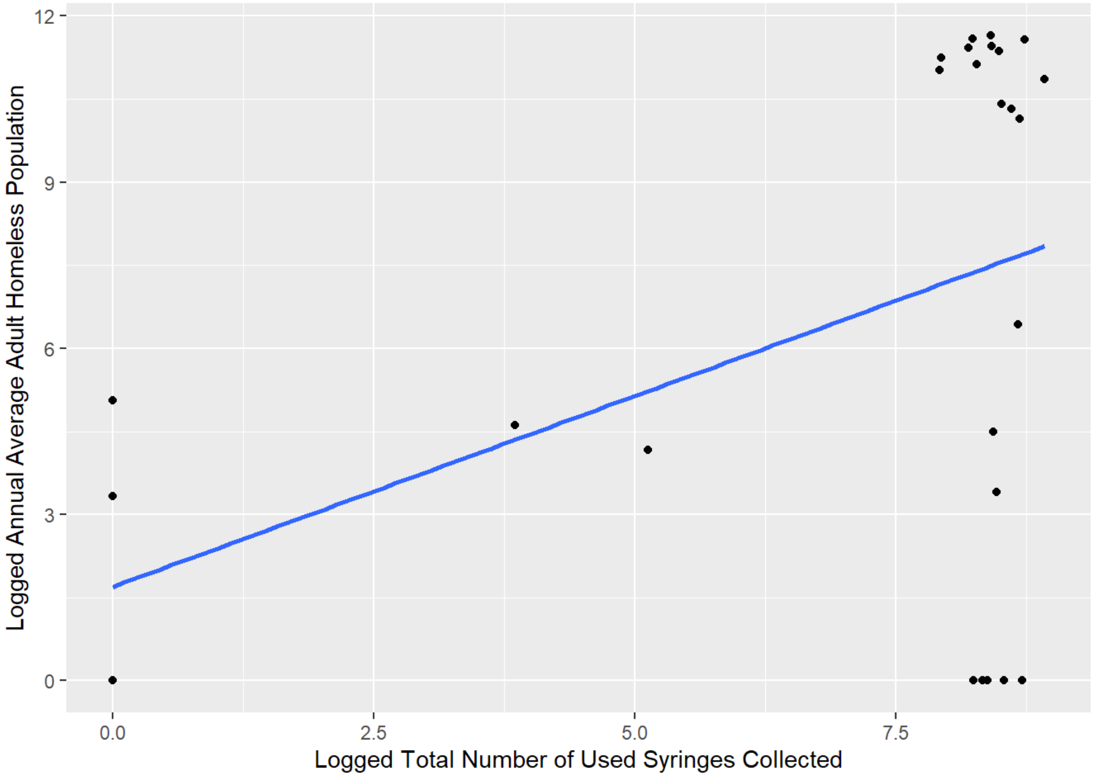


Model 5 may not be Valid. Residuals are mostly normal, but outliers may skew.

QUESTION 2: Does the size of the Adult Homeless Population influence the Number of Used Syringes recovered by NYC Parks?

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

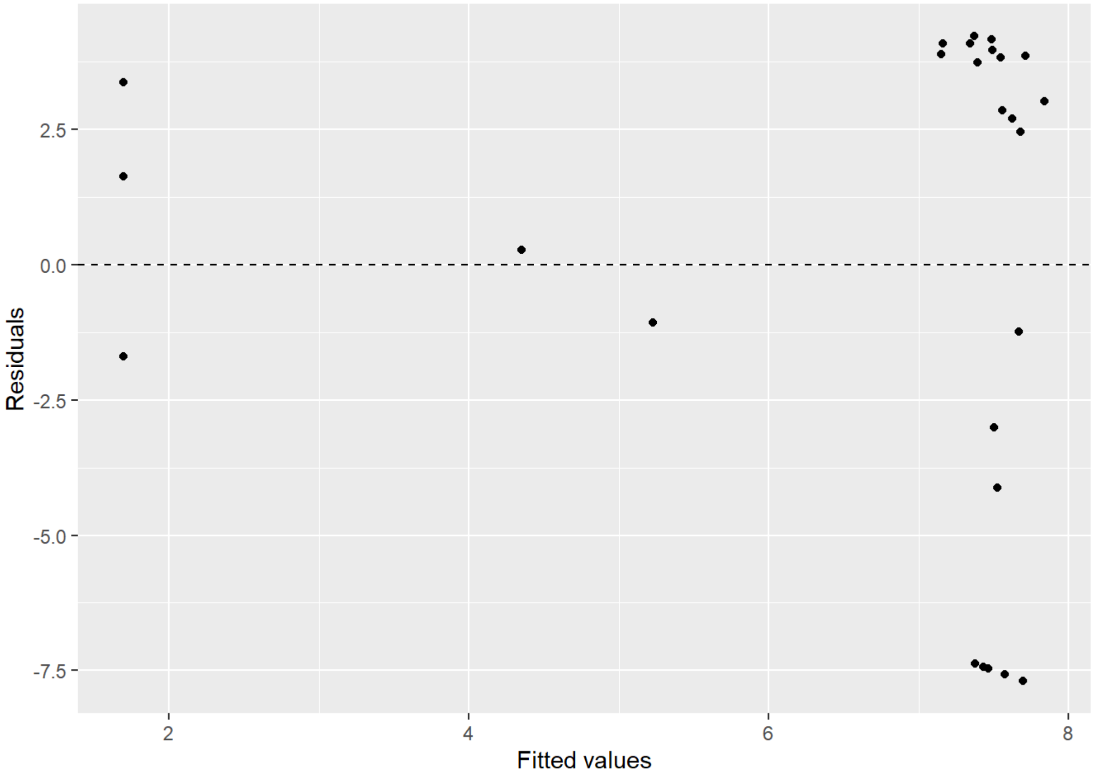
```
## 
## Call:
## lm(formula = log(total_syringes + 1) ~ log(avg_homeless_count + 
##     1), data = data_final_boro_nonulls)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -7.697 -2.024  2.045  3.827  4.218 
## 
## Coefficients:
##                             Estimate Std. Error t value 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
```



Model 6 **is** Statistically Significant. Checking Validity

Model 6 Linearity Check

```
# Linearity Check
ggplot(m6, aes(x=.fitted, y=.resid)) +
  geom_point()+
  geom_hline(yintercept = 0, linetype = "dashed") +
  labs(x = "Fitted values",y ="Residuals")
```

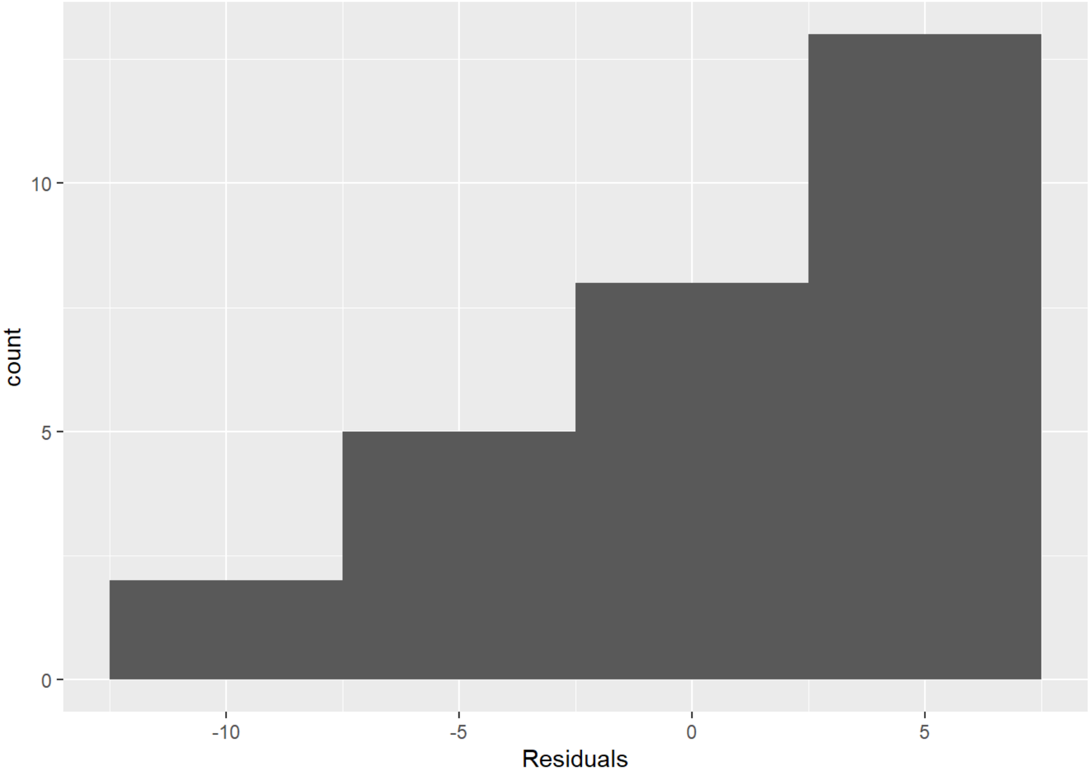


```
##
```

May not be valid, the poitns are clustered and not random.

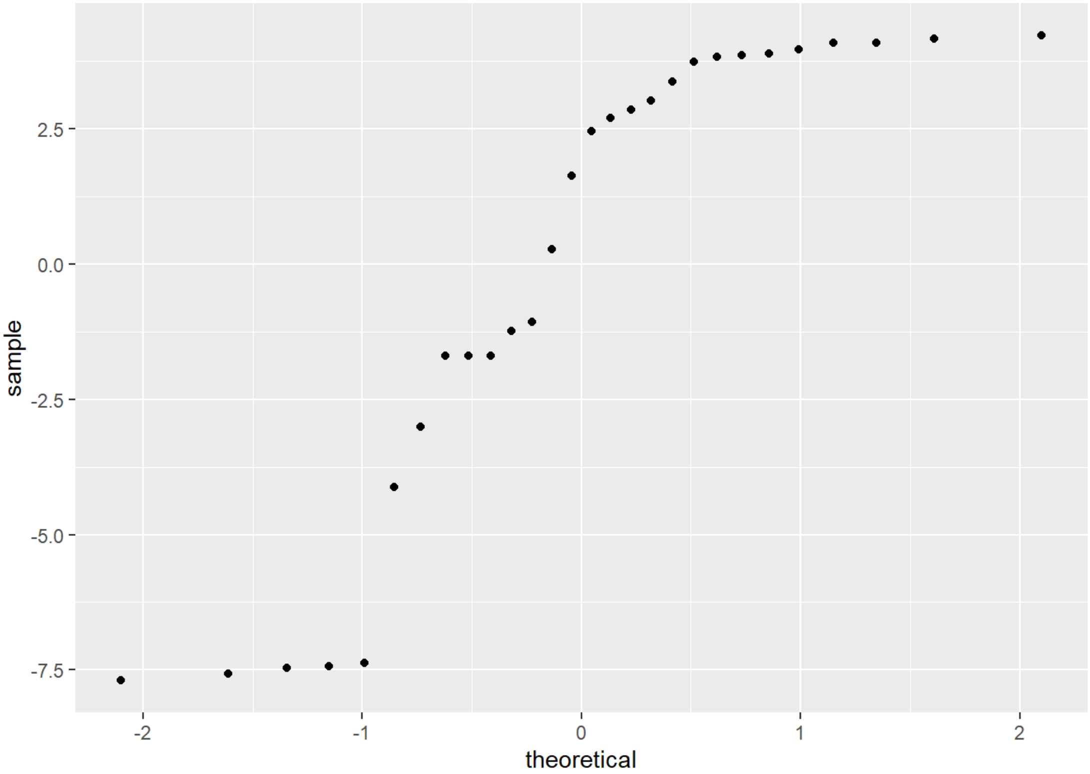
Model 6 Residual Distribution Check 1

```
# Residual Dist.
ggplot(data = m6, aes(x = .resid)) +geom_histogram(binwidth =5) + xlab("Residuals")
```



Model 6 Residual Distribution Check 2

```
#Variablitiy of Constant
ggplot(data = m6, aes(sample = .resid)) + stat_qq()
```



Model 6 may not be Valid. Residuals mostly normal.

Results & Conclusion

Of the 6 different attempts at linear regressions to identify strong relationships between varuables, only 5 had statistically significant results. These results can be seen in the table below. For each of these models, the variable was logged in order to get a more normal distribution.

RESULTS TABLE: Statistically Significant Models

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

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)	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

Community District Level Results

For the analysis at this Geographic level there was two models with a p-value indicating statistical significance. The first was the relationship between the Average Annual Adult Homeless and the total number of Completed Housing related construction projects. About 27% of the variance in the Homeless population can be explained by the Completed Construction Projects. The correlation was direct, which was unexpected. Essentially, the model showed that as the number of completed housing related construction projects for the previous 5 years increased, so did the homeless population. This was surprising because one would expect a negative sloped correlation, or that more construction projects completed the less homeless people. These results refute the Null hypothesis that there is no relationship between the variables, and confirms the alternative hypothesis of the independent variable, completed housing construction projects, impacting the homeless population. However, as mentioned, the result was opposite of what one may expect with a larger homeless population correlating with more completed housing projects in the preceding 5 year window.

The second model that was statistically significant at the Community District level, was the direct relationship between the number of syringes collected in NYC Parks and from public disposal sites and the size of the adult homeless population. Roughly 10% of the variance in Syringe Collection counts can be explained by the size of the adult homeless population. These results reject the null hypothesis and confirm the alternative hypothesis. A reduction in the homeless population does reduce the number of syringes recovered.

Borough Level Results

Due to various limitations of the data when examined at the community district level, the same analysis was carried out at the borough level. This generalization of geographic boundaries can help control for the wide variance of city zoning for shelters, the varying presence of parks, and uneven concentration of housing developments between community districts. All three models at the borough level were statistically significant.

Similar to the Community District analysis, a direct relationship was found between the Adult homeless population and the Total Completed Housing Construction Projects for the previous 5 year window. At the borough level, about 67% of the variance for the size of the homeless population can be explained by the number of completed housing related construction projects for the preceding 5 year window. This is a stronger direct correlation than at the community district level.

Unlike the community district level analysis, the relationship between the YOY average change in completed housing related construction projects and the adult homeless population was statistically significant. This relationship was found to be an indirect one, with about 15% of the variance in the homeless population being explained by the Average YOY Change in projects for the preceding 5 year window. This is more inline with what one would expect, with an increase in housing related construction projects the homeless population would decrease.

Both of these models refute the null hypothesis of no relationship between the completion of housing related construction project and the adult homeless populations, but they do so in different ways.

Lastly, the model at the borough level that explores the relationship between the syringes collected and the size of the adult homeless population shows a direct relationship. An increase in the adult homeless population does increase the number of syringes collected. About 19% of the variance in the amount of syringes collected can be explained by the adult homeless population. This refutes the null hypothesis and confirms the alternative hypothesis for question two.

Overall Conclusion

Overall, the null hypotheses were refuted for both questions, with both alternative hypotheses being confirmed. The number of completed housing related construction projects in the preceding 5 years for an area does in fact correlate to the size of homeless adult shelter population. Similarly, the size of the adult homeless population does correlate with the number of syringes collected from parks and other public disposal units.

The direct relationships between total construction projects completed and homeless population data may be explained by wealthier and gentrifying parts of the city consistently have the attention and resources of developers to create more market-rate and luxury housing via these construction project. The proliferation of such projects may exacerbate the issues surrounding affordable housing. In my view, this is confirmed by the the indirect relationship between the five year average YOY percent change in construction project and the homeless population. Larger YOY increases in completed construction projects may imply a sudden increase in completed construction jobs in areas typically neglected by high-end real estate developers and development general, thus having an indirect relationship with the homeless population. A large average YOY increase in completed construction projects may also imply city planning aimed at helping areas in need of more housing, again helping explain the indirect relationship.

Lastly, the direct relationship between the adult unhoused population and the number of recovered syringes from parks and disposal sites does suggest that this public health issue may be reduced in part by reducing the adult homeless population by increasing the number of housing construction projects.

Limitations of Analysis

While conclusions can be drawn from this analysis there are several data limitations that should be noted. The housing data used in this analysis is for housing-related completed construction projects, this data does not differentiate between high-end developments nor between construction projects that adds housing units to the market, or other types of nuance. More granular housing data would be ideal for this type of regression analysis, such as total units added to market, total affordable housing units added to market, how many rent stabilized units were removed from the market through renovation loopholes, etc.

Further limitations stem from the widely varying variables for each geographic boundary. While Community District variation and the geographic differences in the syringe collection data were accounted for by the aggregation up to the borough level, the variance between boroughs is not accounted for. Each of the boroughs zoning, parks land, and shelter presence varies widely. There was simply not enough years of data to analyze these variables at the city-level.

Additionally, the adult homeless population data used in this analysis originated at counts from the shelter system. This does not take into account the unhoused population that makes no use of the shelter system, so there may be a large blind spot not considered in this analysis.

Lastly, several of the variables had floor effects on their distributions due to the nature of the variables. Log transformations were used, but there still were a large number of zero values for several of the variables examined. This analysis, other than logging the data, did not accommodate for this by removing outliers, weighing data points, or by other means. With this said, as shown by the model validity tests for linearity, some patterns are discernible in the data. This may mean that these models are invalid and the conditions for linearity may not be met. The inclusion of these zero values and some outliers is another limitation of this analysis. Future work, should attempt to further clean and process the data so as to create more valid models.

References

[1] “Home Prices, Wages, and Rent: Harvard Report Reveals Housing Trends.” NPR, 20 June 2024, https://www.npr.org/2024/06/20/nx-s1-5005972/home-prices-wages-paychecks-rent-housing-harvard-report#:~:text=In%20past%20decades%2C%20it%20was,Philly%20is%20giving%20renters%20cash (). Accessed 7 Dec. 2024.

In addition to this citation, various internet sources like StackOverFlow and other blogs were referenced for R syntax.