# Appendix 2: RIDP Serial Evictors Data Cleaning and Analysis

Angel Aliseda - GovEx Research & Analytics

2025-02-20

## Contents

## Read data

```
i_am("allentown_ridp_evictors_analysis.Rmd")

## Read list of evictions in Allentown from previously cleaned evictions data
allentown_evictions_df <- read_csv(here("output", "allentown_evictions_full_df.csv"))

## Deduplicate judgment components to obtain unique cases and the plaintiff for each case
allentown_evictors_df <- allentown_evictions_df |>
  mutate(plaintiff_name = replace_na(plaintiff_name, "Not Available")) |>
  group_by(docket_number, plaintiff_name) |>
  summarise(plaintiff_name = unique(plaintiff_name))

total_plaintiffs_before_cleaning <- allentown_evictors_df |>
  group_by(plaintiff_name) |>
  summarise(total = n()) |>
  nrow()

## Obtain total number of eviction cases
total_evictions_allentown <- allentown_evictions_df |>
  select(docket_number) |>
  summarise(filings = unique(docket_number)) |>
  summarise(total_filings = n()) |>
  pull(total_filings)
```

Initially, before any cleaning and standardizing, the dataset contained 1665 unique plaintiffs. However, after an initial screening we realized there are errors in the spelling of some plaintiff's names that need to be reviewed.

## Standardize Format

```
## Standardize text formatting among plaintiffs
allentown_evictors_formatted_df <- allentown_evictors_df |>
  mutate(plaintiff_name = str_to_lower(plaintiff_name),
         plaintiff_name = str_squish(plaintiff_name)) |>
  mutate(plaintiff_name = str_remove_all(plaintiff_name, "[.,/\\\\]"),
         plaintiff_name = str_replace_all(plaintiff_name, "west", "w"),
         plaintiff_name = str_replace_all(plaintiff_name, "east", "e"),
         plaintiff_name = str_replace_all(plaintiff_name, "north", "n"),
         plaintiff_name = str_replace_all(plaintiff_name, "south", "s"),
         plaintiff_name = str_replace_all(plaintiff_name, "street", "st"),
         plaintiff_name = str_replace_all(plaintiff_name, "mgmt", "management"),
         plaintiff_name = str_replace_all(plaintiff_name, "mgt", "management"))

## Create a table with the list of unique evictors after initial text formatting
allentown_unique_evictors_formatted_df <- allentown_evictors_formatted_df |>
  group_by(plaintiff_name) |>
  summarise(total = n(),
            percentage = total/total_evictions_allentown*100)
```

First, we need to standardize how the names are written. We need to:

- Convert all names to lower case
- Remove special characters such as "." "," "/" and ""
- Standardize how street names are written, for example change "west" to "w" and "street" to "st"
- Standardize how specific words are being abbreviated such as using "management" instead of "mgmt" or "mgt".

After this format standardization, we were able to reduce the number of unique values to 1567

## Levenshtein Distance Matching Methodology

```
## Obtain the list of unique landlords after text formatting
#allentown_evictors_unique <- unique(allentown_evictors_formatted_df$plaintiff_name)

## Calculate the Levenshtein distance among all the different plaintiffs
### The Levenshtein distance measures the difference between two strings by
### counting the minimum number of single-character edits (insertions, deletions,
### or substitutions) required to transform one string into another. It's useful
```

```r
### for similarity matching, such as comparing names with slight spelling differences.

lv_matrix <- stringdistmatrix(allentown_unique_evictors_formatted_df$plaintiff_name,
                              allentown_unique_evictors_formatted_df$plaintiff_name,
                              method = "lv")

## Replace row and column names for clarity
rownames(lv_matrix) <- allentown_unique_evictors_formatted_df$plaintiff_name
colnames(lv_matrix) <- allentown_unique_evictors_formatted_df$plaintiff_name

## Set threshold for similarity
threshold <- 3

# Find closest match for each plaintiff name within the threshold
matches <- apply(lv_matrix, 1, function(row) {
  min_dist <- min(row[row > 0])  # Ignore self-matches (distance = 0)
  if (min_dist <= threshold) {
    closest_match <- colnames(lv_matrix)[which.min(row[row > 0])]
    return(closest_match)
  } else {
    return(NA)  # No match if distance exceeds threshold
  }
})

# Combine results into a data frame
lv_matches_evictors <- data.frame(
  Plaintiff_Name = allentown_unique_evictors_formatted_df$plaintiff_name,
  Closest_Match = matches) |>
  filter(!is.na(Closest_Match))
```

Then use the Levenshtein Distance (LV) methodology to find similar names across the dataset. LV Distance measures the difference between two strings by counting the minimum number of single-character edits (insertions, deletions, or substitutions) required to transform one string into another. It's useful for similarity matching, such as comparing names with slight spelling differences. For this analysis, we looked at pairs with a similarity (or LV distance) of 3 or less.

## Manual Matches after LV Matching and ChatGPT

After obtaining the LV pairs that met the threshold, we manually reviewed and edited the corresponding names if we agreed they belonged to the same plaintiff. Additionally, we used ChatGPT to find other similar names that might not have been paired using the LV Distance methodology using the following prompt:

- "Review the following list of entities and make suggestions of similar names that could refer to the same entity but might have differences in spelling or formatting. These list contains a list of plaintiffs from eviction case filings. These names were taken directly from court records. Some of them are LLCs or other companies and some are individual landlords."

```r
# After manually checking the closes match for each evictor in the lv matrix and
# using ChatGPT, manually edit the plaintiff names that should be matched
## Matching plaintiff whose name is an address and are missing "st" after
## the street name
## Matching plaintiffs with missing cardinal point (n, s, w, or e) in the name
## and there is another plaintiff with the exact same address but with the proper
## cardinal point
## Matching plaintiffs with missing or extra spaces or individual strings
allentown_evictors_formatted_matched_df <- allentown_evictors_formatted_df |>
  mutate(plaintiff_name = case_match(plaintiff_name,
                                     "1039 mechanic llc" ~ "1039 mechanic st llc",
                                     "142 n 7th llc" ~ "142 n 7th st llc",
                                     "1550 warren st allentown llc allentown pa" ~ "1550 warren st allen
                                     "401 n 8th st llc co" ~ "401 8th st llc",
                                     "440 tilghman st llc" ~ "440 w tilghman st llc",
                                     "501 n front llc" ~ "501 n front st llc",
                                     "520 hamilton oplp" ~ "520 hamilton op lp",
                                     "626-628 n 6th llc" ~ "626-628 n 6th st llc",
                                     c("631-645 n jordan llc",
                                       "631-645 n jordan st llc allentown") ~ "631-645 n jordan st llc"
                                     "727 n meadow st llc allentown" ~ "727 n meadow st llc",
                                     c("734 n railroad llc",
                                       "734 railroad st llc") ~ "734 n railroad st llc",
                                     c("900 hamilton st associates lp",
                                       "900 hamilton st assoc lp") ~ "900 hamilton st associates",
                                     "941 w hamilton corp llc nj" ~ "941 w hamilton corp llc",
                                     "abc realty ll holdings llc" ~ "abc realty ii holdings llc",
                                     "abdouche george s" ~ "abdouche george",
                                     c("ag investments llc allentown",
                                       "ag investment llc") ~ "ag investments llc",
                                     "allentown housing authority allentown" ~ "allentown housing autho
                                     c("allentown metro holding lp",
                                       "allentown metro holding") ~ "allentown metro holdings",
                                     "alonzo roosvelt" ~ "alonzo roosevelt",
                                     "anb realty 5 llc" ~ "anb realty 5",
                                     "anthony howard lp allentown" ~ "anthony howard lp",
                                     c("arayus management allentown",
                                       "arayus managment",
                                       "arayvs managment") ~ "arayus management",
                                     c("arch ventures llc co del vel realty & proprty mgmt",
                                       "arch ventures llc co del vel realty & proprty management") ~ "a
                                     "armour court" ~ "armour court llc",
                                     "baechle mark t" ~ "baechle mark",
                                     "baez jaeinto" ~ "baez jacinto",
                                     "attieh tony m" ~ "attieh tony",
                                     "batista johnney w ii" ~ "batista johnny w ii",
                                     "big m holding co" ~ "big m holding company",
                                     "bucks dean" ~ "buck dean",
                                     "big mom inc allentown" ~ "big mom inc",
                                     "blanco realty" ~ "blanco realty inc",
                                     "bluelakes llc allentown" ~ "bluelakes llc",
                                     c("campbell real estate investment",
                                       "campbell real estate investments",
                                       "campbell real estate investments llc bethlehem",
                                       "campbell real estate invest llc bethlehem",
                                       "campbell real estate investments bethlehem") ~ "campbell real es
                                     c("capece dave s",
                                       "capece david s",
                                       "capece david a") ~ "capece david",
                                     "casmre g develop" ~ "casmre g development",
                                     c("cdc developers inc allentown",
                                       "cdc develpoers in") ~ "cdc developers inc",
                                     "center square lofts e op lp" ~ "center square lofts op lp",
                                     "centerpool llc" ~ "centerpol llc",
                                     "chalaby salem" ~ "chaleby salem",
                                     "cruz manuel a" ~ "cruz manuel",
                                     c("dang phuong (peter) t",
                                       "dang peter") ~ "dang phuong (peter)",
                                     "dayoub micheal" ~ "dayoub michael",
                                     "decker and clark llc" ~ "decker & clark llc",
                                     "de falco paul" ~ "defalco paul",
                                     "diaz nelson a" ~ "diaz nelson",
                                     "dnd realty holdings llc allentown" ~ "dnd realty holdings llc",
                                     "dnj properties" ~ "dnj properties llc",
                                     c("dream big investment llc",
                                       "dream big investments 1030 w hamilton st",
                                       "dream big investments management llc") ~ "dream big investments
                                     c("driscoll tim f",
                                       "driscoll tim") ~ "driscoll timothy",
                                     "dt captial investments llc" ~ "dt capital investments llc",
                                     "dunn daniele" ~ "dunn danielle",
                                     c("e coast whispering hills llc tdba lehigh square apartments",
                                       "e coast whispering hills llc tdba whispering hills",
                                       "e coast whispering hillsllc") ~ "e coast whispering hills llc",
                                     "e&b real estate holdings penn #1" ~ "e&b real estate allentown pa
                                     "emd 14 llc" ~ "edm 14 llc",
                                     "fieldstone associates lp allentown" ~ "fieldstone associates lp th
                                     "g & b allentown properties" ~ "g&b allentown properties",
                                     "g & b eon properties" ~ "g&b eon properties",
                                     "gomez yessenia" ~ "gomez yesenia",
                                     "grewal jagjeet s" ~ "grewal jagjeet",
                                     "haas nina" ~ "haas ning",
                                     "hamilton gardens apartments" ~ "hamilton gardens allentown",
                                     "hamilton towers" ~ "hamilton tower",
                                     "hillegas stephen t" ~ "hillegas stephen",
                                     "home investment" ~ "home investment llc",
                                     "home snipe real estate llc allentown" ~ "homesnipe real estate llc
                                     "hoffman eric p" ~ "hoffman eric",
                                     "hotel traylor llc allentown" ~ "hotel traylor llc",
                                     "howard lp anthony" ~ "howard anthony",
                                     "hp altman hidden village apts allentown" ~ "hp altman hidden villa
                                     "hq partners bethlehem" ~ "hq partners llc bethlehem",
                                     "ideal management group llc" ~ "ideal management group",
                                     "is capital investment llc"  ~ "is capital investments llc",
                                     "jat construction development" ~ "jat construction development llc
                                     "jet-set enterprises llc attn: gerry sanchez" ~ "jet-set enterprise
                                     "jsdsjs1 llc" ~ "jsdsjs 1 llc",
                                     "kheir & kellar investment partners llc" ~ "kheir & kellar investme
                                     "laua khemraj" ~ "lalla khemraj",
                                     "lavigne enterprises llc" ~ "lavigne enterprises",
                                     "lazarus housing llc" ~ "lazarus house llc",
                                     c("leftkiki 510 llc",
                                       "leftykiki 510") ~ "leftykiki llc",
                                     c("lehigh land holdings inc",
                                       "lehigh landholdings iinc",
                                       "lehigh landholding",
                                       "lehigh landholdings inc allentown") ~ "lehigh landholdings inc"
                                     "lehigh valley management & solutions llc" ~ "lehigh valley manage
                                     c("lehigh valley private equity fund",
                                       "lehigh valley private equity fund co") ~
                                       "lehigh valley private equity fund llc",
                                     "linden st commons op llp" ~ "linden st commons op lp",
                                     c("loth investments co",
                                       "loth investments company") ~ "loth investments",
                                     "m & m leon inc allentown" ~ "m&m leon inc",
                                     "mahmoud omar m" ~ "mahmoud omar",
                                     "makhoul electric" ~ "makhoul electric llc",
                                     "malhotra harden"  ~ "malhotra hardev",
                                     "matthew + margaret ricchio" ~ "matthew & margaret ricchio",
                                     "matos henrique (rick)" ~ "matos henrique",
                                     "minnas balji" ~ "minhas balji",
                                     "morales amelty" ~ "morales amelfy",
                                     "morocho investments llc allentown pa" ~ "morocho investments llc"
                                     "namous ali a" ~ "namous ali",
                                     "nova lehigh holding company llc newark" ~ "nova lehigh holding co
                                     "nunez frankeirys r" ~ "nunez frankeirys",
                                     "nunez santos adriano a" ~ "nunez santos adriano",
                                     "ojong peter b" ~ "ojong peter",
                                     "old forge apartments" ~ "old forge apts",
                                     "oomen john" ~ "oommen john",
                                     "pa i hamilton 1129 llc" ~ "pa 1 hamilton 1129 llc",
                                     "park hyongjoon" ~ "park hyoungjoon",
                                     "penn crest associates" ~ "penn crest associates allentown",
                                     "placencio francisco" ~ "plasencia francisco",
                                     "pinnam reddy nithin reddy" ~ "pinnam reddy yugandh nithin reddy",
                                     "rennig cory" ~ "rennig corey",
                                     "regency towers apartments" ~ "regency towers opco llc tdba regency
                                     "riverview lofts allentown" ~ "riverview lofts allentown llc",
                                     "rivera jose a" ~ "rivera jose",
                                     "rlp2 llc" ~ "rlp 2 llc",
                                     "roba michael e" ~ "roba michael",
                                     "royal lika properties allentown" ~ "royal lika properties",
                                     "said mike a" ~ "said mike",
                                     "saltz keren" ~ "saltz karen",
                                     "seabear real estate llc & next summit llc" ~ "seabear real estate
                                     "society hill apartment allentown" ~ "society hill apartments allen
                                     "starbrite realty bethlehem" ~ "starbright realty bethlehem",
                                     "third properties llc" ~ "thind properties llc",
                                     "three m holding company llc" ~ "three m holding company",
                                     "toomey investment group llc" ~ "toomey investment group",
                                     "turner apartments" ~ "turner st apartments",
                                     "tracker investment group llc" ~ "tracker investment group",
                                     "twelfth cola llc" ~ "twelfth cola",
                                     "vazquez123 realty llc" ~ "vazquez 123 realty llc",
                                     "walnut st commons l lp" ~ "walnut st commons i lp",
                                     c("wikabayashi rikako",
                                       "wakabayashi sen ando rikako") ~ "wakabayashi rikako",
                                     "waldemar lichmira carol delancey samuel lichmira empire property
                                     "whitestone village apt assoc llc allentown" ~ "whitestone village
                                    "william dawson llc co empire management group llc allentown pa" ~
                                     "wynnewood greens allentown" ~ "wynnewood greens",
                                     "yda properties llc co" ~ "yda properties llc",
                                    "zre services llc helen muniz catasauqua"  ~ "zre services llc-helen
                                     .default = plaintiff_name),
         plaintiff_name = case_when(grepl("depaul m", plaintiff_name) ~
                                      "depaul management company",
                                    grepl("equinox property management", plaintiff_name) ~
                                      "equinox property management",
                                    grepl("jb enterprise", plaintiff_name) ~ "jb enterprise of pa",
                                    grepl("jmd property", plaintiff_name) ~
                                      "jmd property investments inc",
                                    grepl("kellar properties", plaintiff_name) ~
                                      "kellar properties llc",
                                    grepl("full circle realty & property", plaintiff_name) ~
                                      "full circle realty & property management",
                                    grepl("hudson homes management llc as attorney in fact for us bank
                                      "hudson homes management llc as attorney in fact for us bank trus
                                    grepl("mk barreto real estate", plaintiff_name) ~
                                      "mk barreto real estate holding llc",
                                    grepl("riverbend", plaintiff_name) ~ "riverbend in allentown",
                                    grepl("park run ", plaintiff_name) ~
                                      "park run management",
                                    grepl("safe home ", plaintiff_name) ~ "safe home investment corp",
                                    grepl("smith daniel", plaintiff_name) ~ "smith daniel",
                                    grepl("scully company", plaintiff_name) ~
                                      "scully company dba bridgeview apartments",
                                    grepl("sterling equity group", plaintiff_name) ~ "sterling equity g
                                    grepl("willow matthew", plaintiff_name) ~
                                      "willow matthews llc",
                                    .default = plaintiff_name))
```

Rules and decisions made to match similar names

## Plaintiff Names that were matched

To decide whether two plaintiff names were the same, I followed these rules for the entire list of suggested pairings with a Levenshtein distance of 3 or less. I also used the spelling with most occurrances in the original dataset.

- Matched plaintiff whose name is an address and are missing “st” after the street name
    - For example: “1039 mechanic llc” was matched with “1039 mechanic st llc”
- Matched plaintiffs with missing cardinal point (n, s, w, or e) in the name if there is another plaintiff with the exact same address but with a cardinal point
    - For example: “440 tilghman st llc” was matched with “440 w tilghman st llc”

- Matching plaintiffs with extra strings
  - For example: “baechle mark t” was matched with “baechle mark”
- Matching plaintiffs with missing or extra spaces
  - For example: “520 hamilton oplp” was matched with “520 hamilton op lp”

Specific cases:

- Plaintiffs “capece dave s”, “capece david a”, “capece david s”, and “capece david” were considered the same plaintiff.
- Plaintiffs “dang phuong (peter) t”, “dang phuong (peter)”, and “dang peter” were considered the same plaintiff.
- All plaintiffs that had “depaul management company” in the name were considered the same plaintiff
- All plaintiffs that had “equinox property management” in the name were considered the same plaintiff.
- “lazarus housing llc” and “lazarus house llc” were considered the same plaintiff
- “e coast whispering hillsllc”, “e coast whispering hills llc tdba lehigh square apartments”, and “e coast whispering hills llc tdba whispering hills” were considered the same plaintiff but they were not matched with “e coast whispering hills llc tdba whispering hills co cohen marraccini llc” because of the additional co-plaintiff
- All plaintiffs with “riverbend” in the name were matched as “riverbend in allentown”

### Plaintiff Names that were NOT matched

However, there were some plaintiffs that even though are really similar, I decided not to match because I couldn’t be sure if they were the same entity. For example:

- Plaintiffs mentioned on filings with other co-plaintiffs.
  - For example: “empire property management group llc” was not matched with “matt donely co empire property management group llc”
- Plaintiffs that were individuals with the same name but with a different suffix.
  - For example: “bauer robert” was not matched with “bauer robert jr”
- Companies that had the same name but add an extra number at the end.
  - For example: “ag investment llc” was not matched with “ag investments 2 llc” or with “ag investments 3 llc”
- g & b eon properties and g & b allentown properties were considered different plaintiffs
- zfp009 llc and zfp007 llc were considered different plaintiffs

## Top Evictors Graph

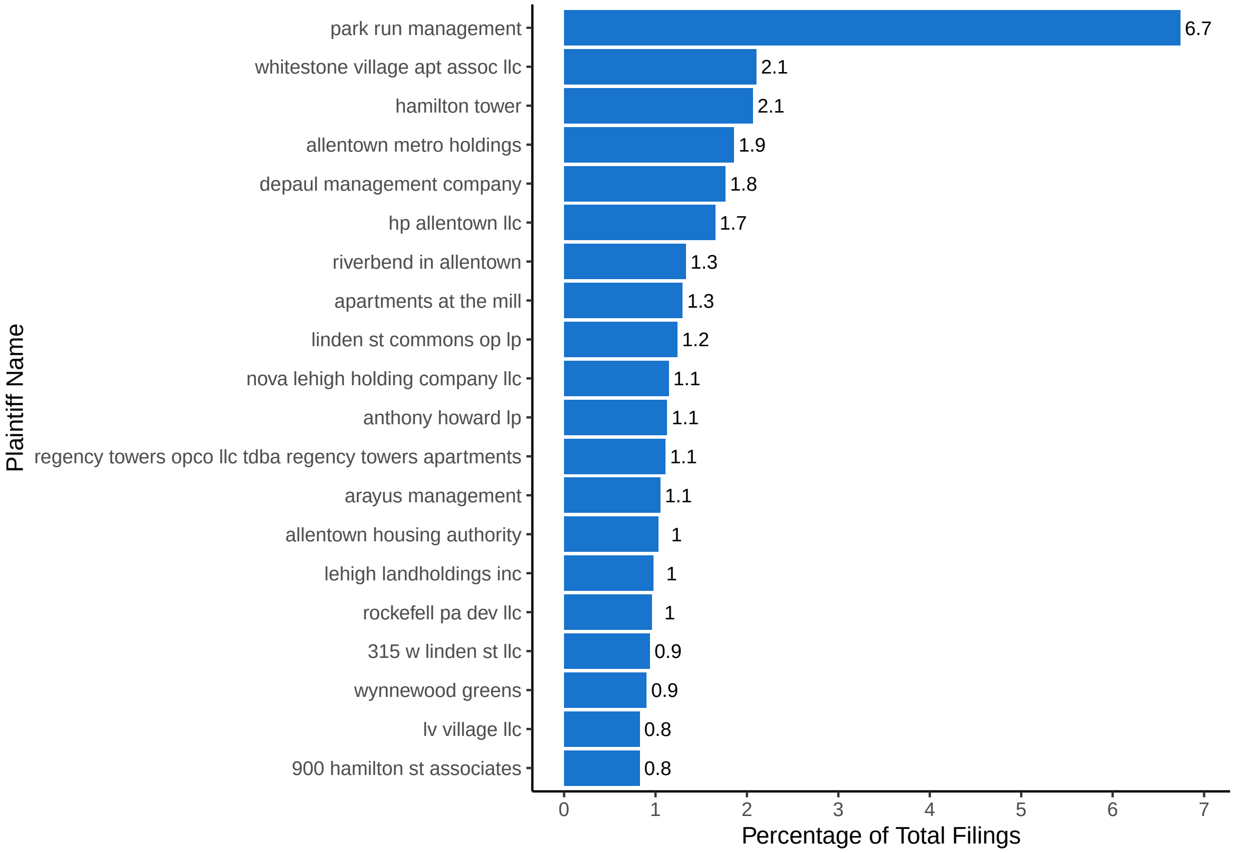
```r
## summarize table to obtain the total number of evictions by unique evictor/plaintiff after text forma
allentown_evictors_totals <- allentown_evictors_formatted_matched_df |>
  group_by(plaintiff_name) |>
  summarise(total = n(),
            percentage = total/total_evictions_allentown*100)

## Save table as csv
write_csv(allentown_evictors_totals,
          here("output", "allentown_evictors_formatted_matched_totals.csv"))
```

After the manual match, we finalized with 1343 unique plaintiffs. The 20 plaintiffs that appeared the most in the dataset are represented in the following graph

```r
## Plot of top evictors in Allentown after cleaning, standardizing formatting
## and LV matching
allentown_evictors_totals |>
  arrange(desc(percentage)) |>
  head(20) |>
  ggplot(aes(x=reorder(plaintiff_name, percentage), y = percentage)) +
  geom_col(fill = "dodgerblue3") +
  geom_text(aes(label= round(percentage,1)), nudge_y = .2, size= 3) +
  coord_flip() +
  labs(title = "Figure A2.1. Plaintiffs with Most Eviction Filings in Allentown ",
       x = "Plaintiff Name" ,
       y = "Percentage of Total Filings",
       caption = "Source: Own elaboration with data from the Housing Alliance of Pennsylvania.") +
  scale_y_continuous(breaks = seq(0,7,1))+
  theme_classic()
```

Figure A2.1. Plaintiffs with Most Eviction Filings in Allento



Source: Own elaboration with data from the Housing Alliance of Pennsylvania.