

Analysis Component Group E

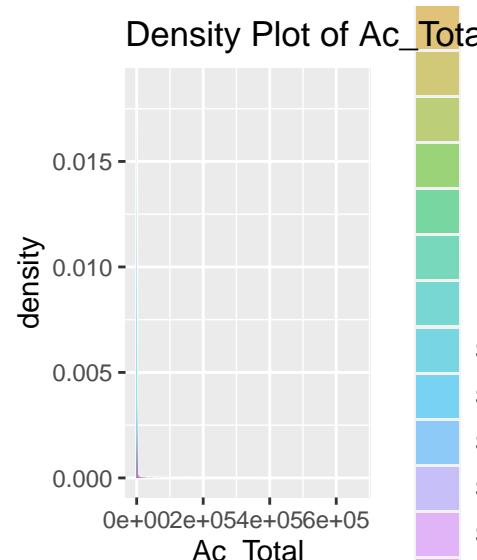
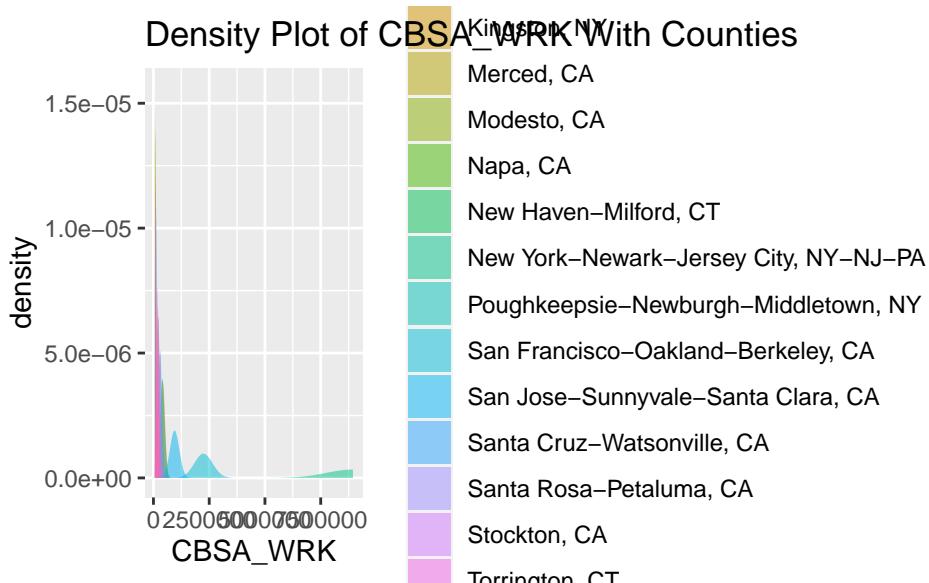
E. Hauge, G. Henry, L. Kaminka

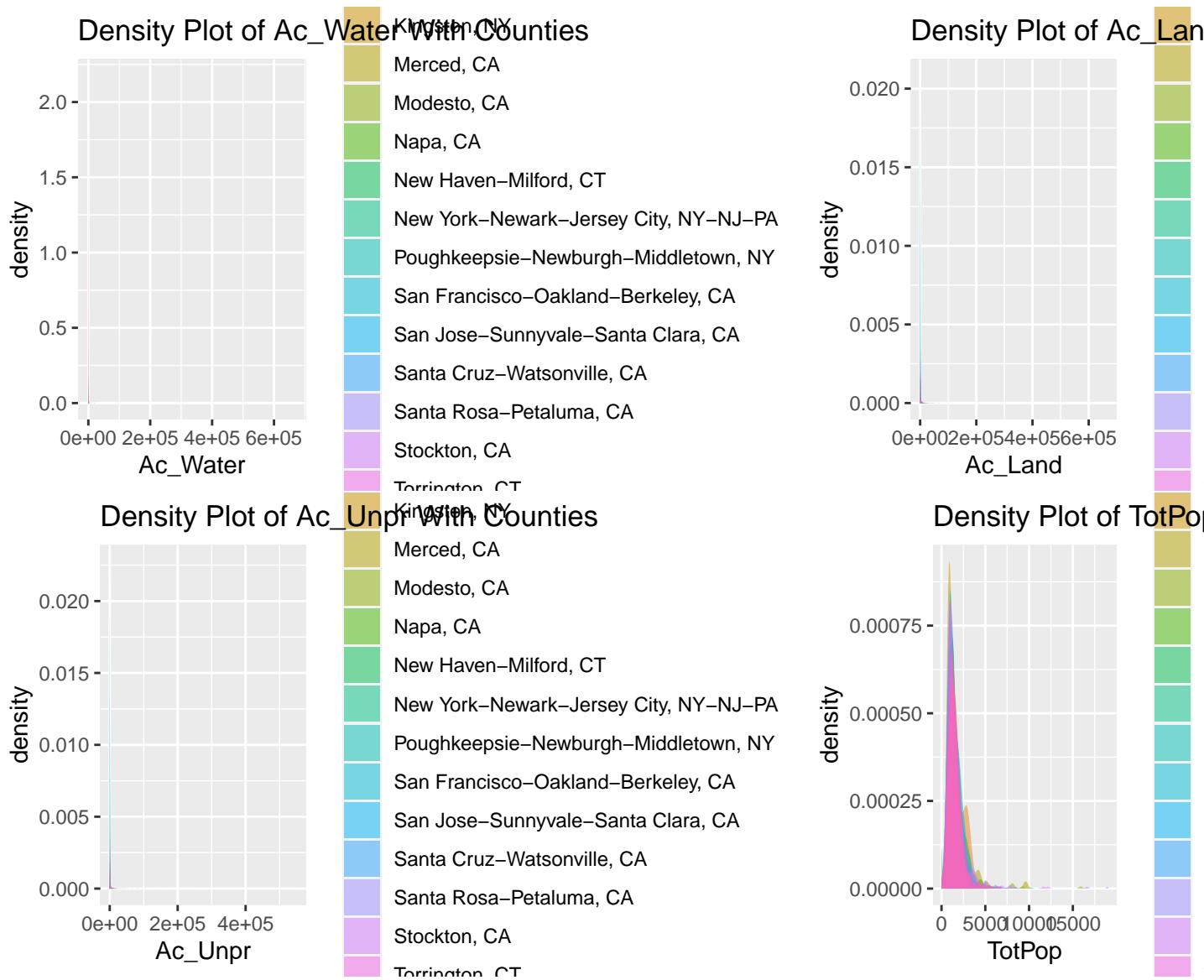
2024-03-28

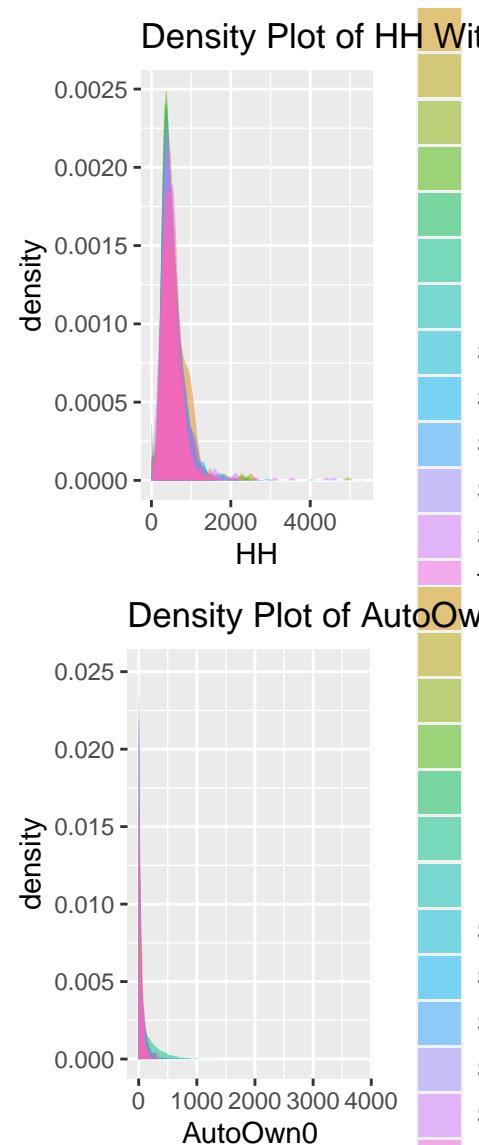
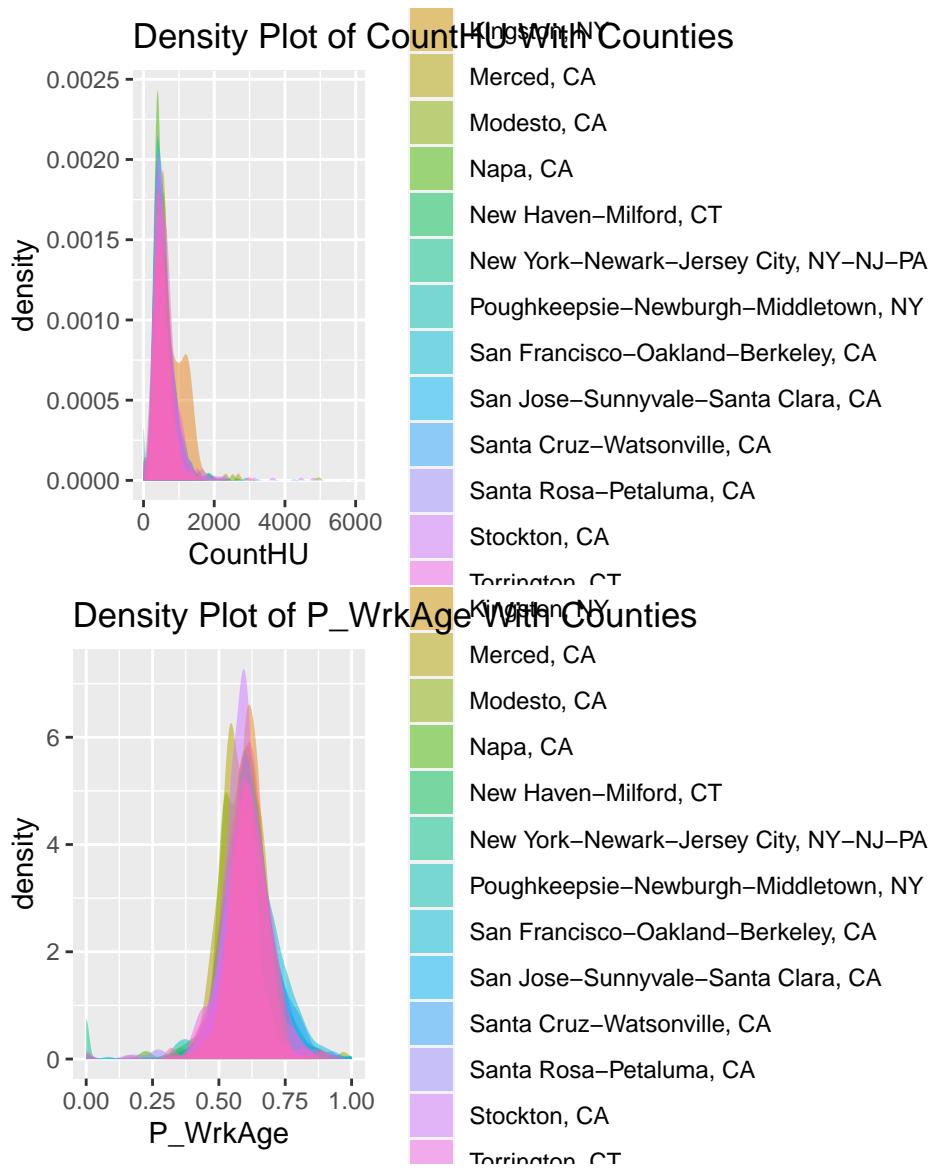
```
SLD <- read_csv("data/NewYork_SanFrancisco_data.csv", show_col_types = FALSE)
```

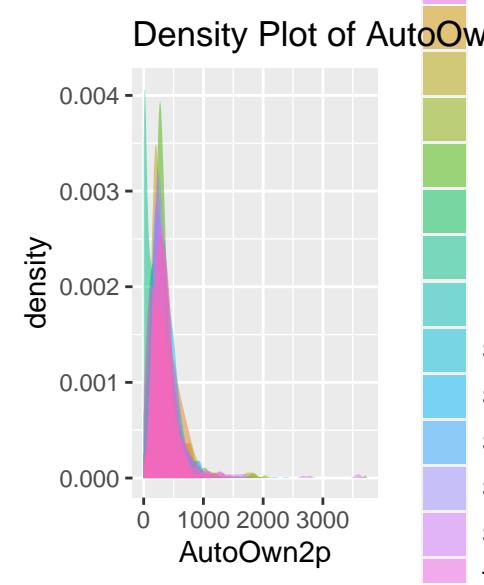
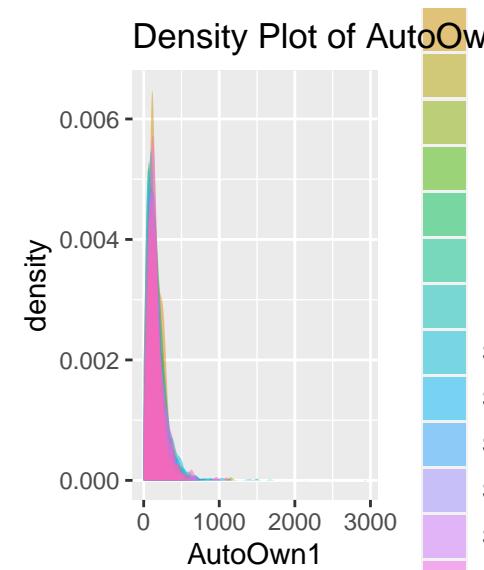
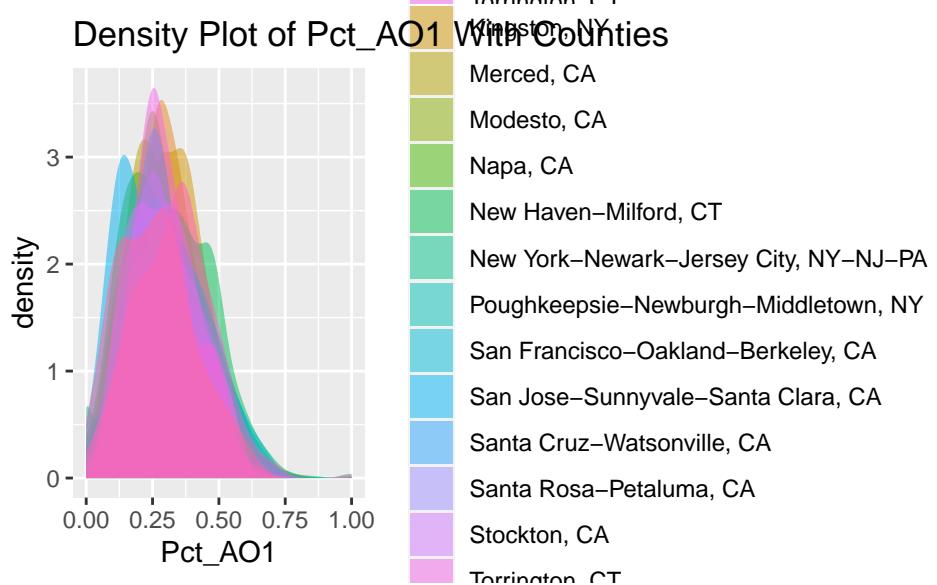
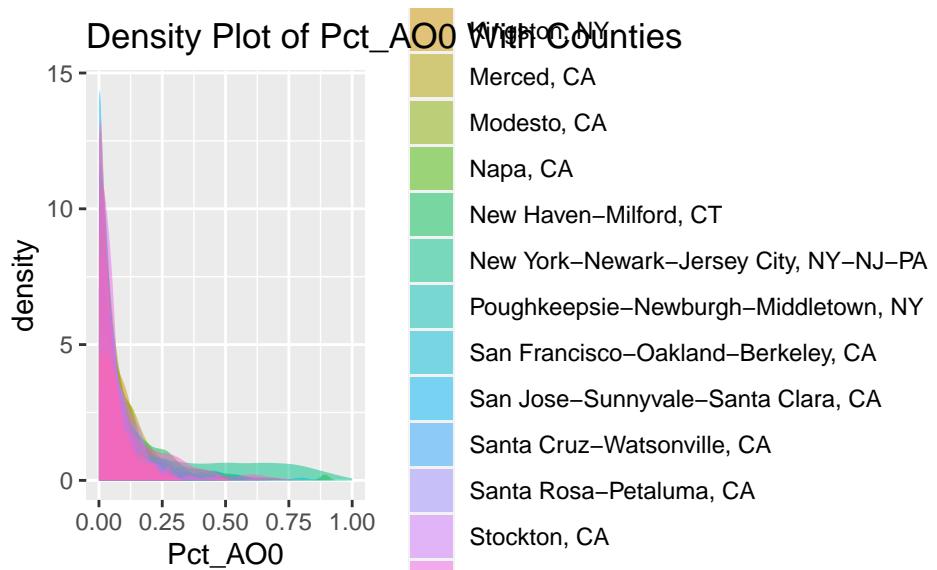
```
## New names:  
## * ` ` -> '...1'  
  
create_density_plots_all <- function(data, start_col, end_col) {  
  numeric_vars <- names(data)[sapply(data, is.numeric)]  
  selected_vars <- numeric_vars[start_col:end_col]  
  
  for (var in selected_vars) {  
    # Create density plot for each variable  
    plot <- gf_density(as.formula(paste("~", var)), data = data, fill = ~CBSA_Name,  
                       title = paste("Density Plot of", var, "With Counties"))  
    print(plot)  
  }  
}
```

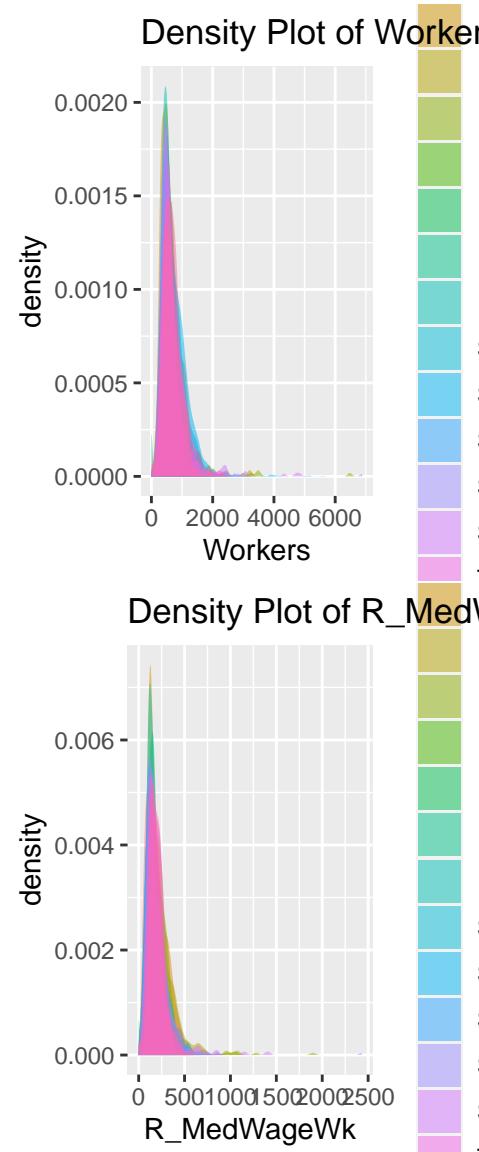
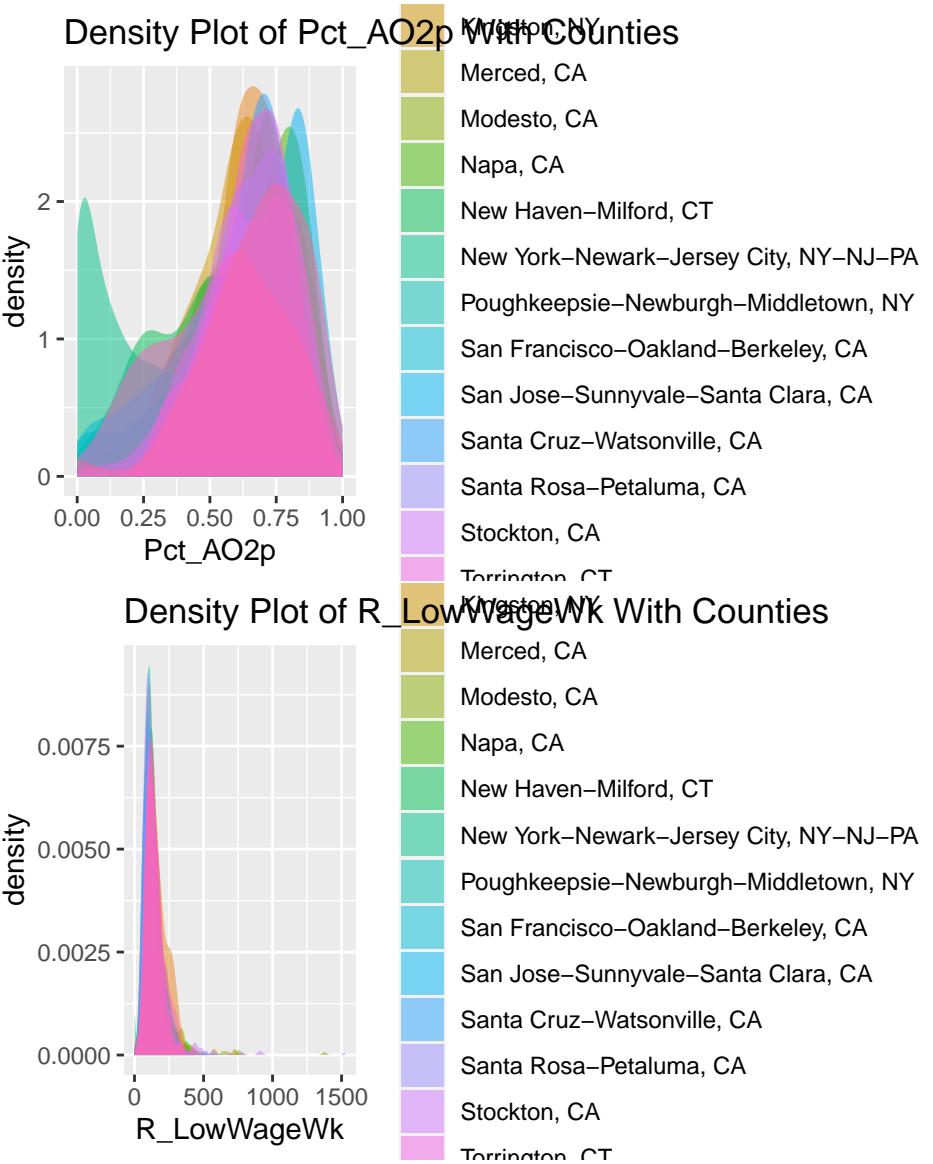
```
create_density_plots_all(SLD, 13, 33)
```

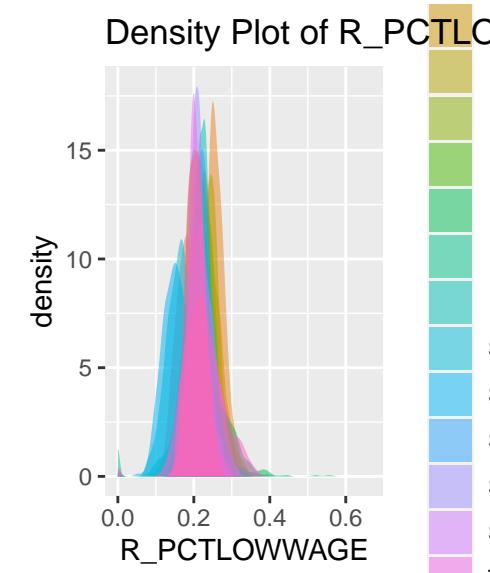
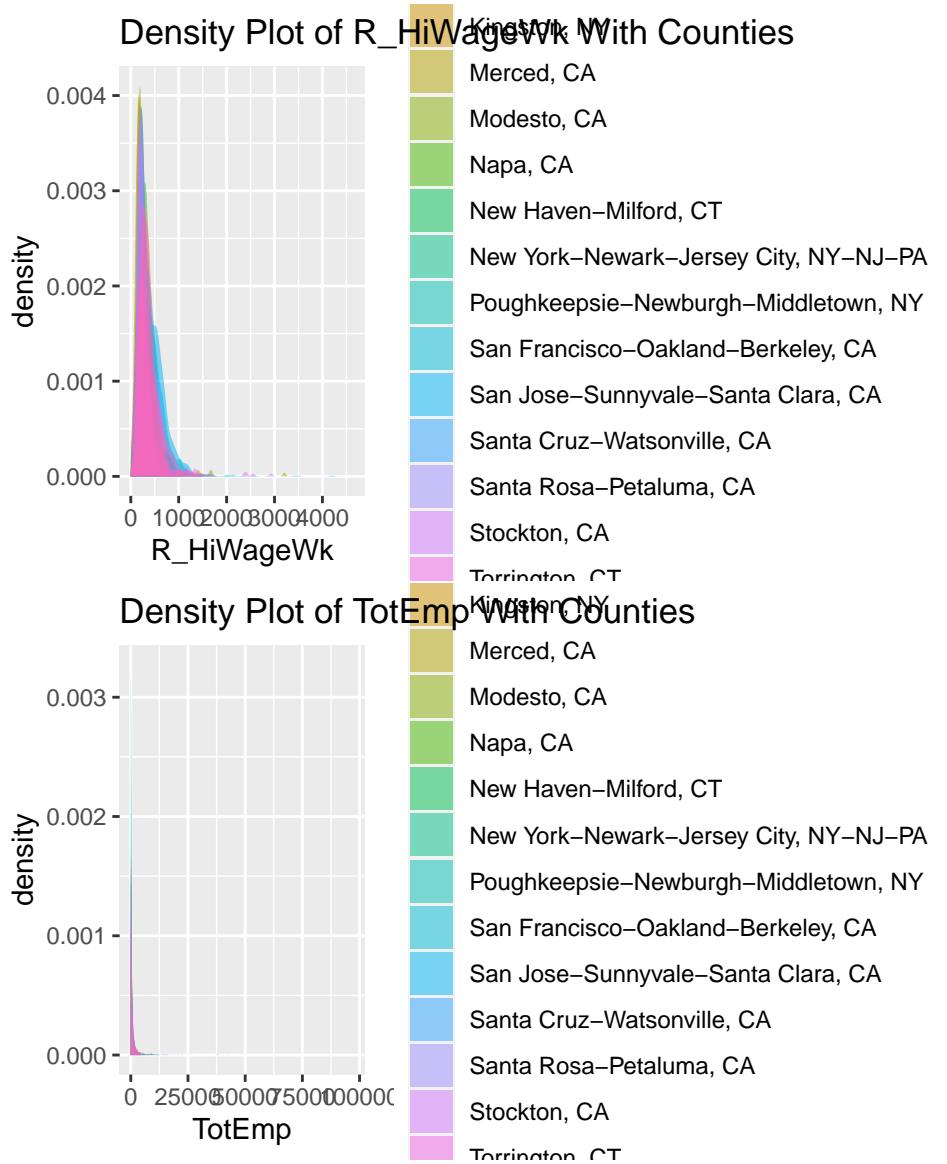






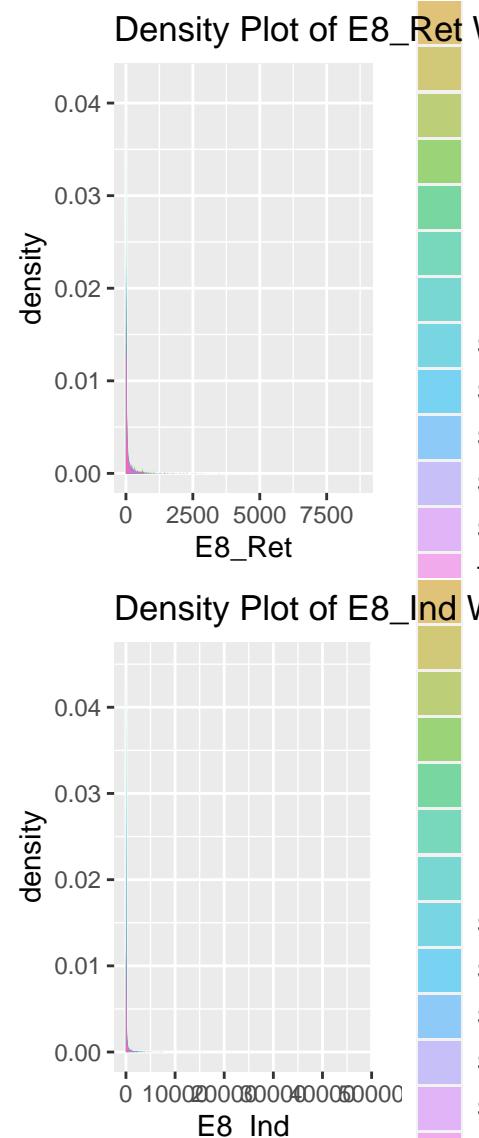
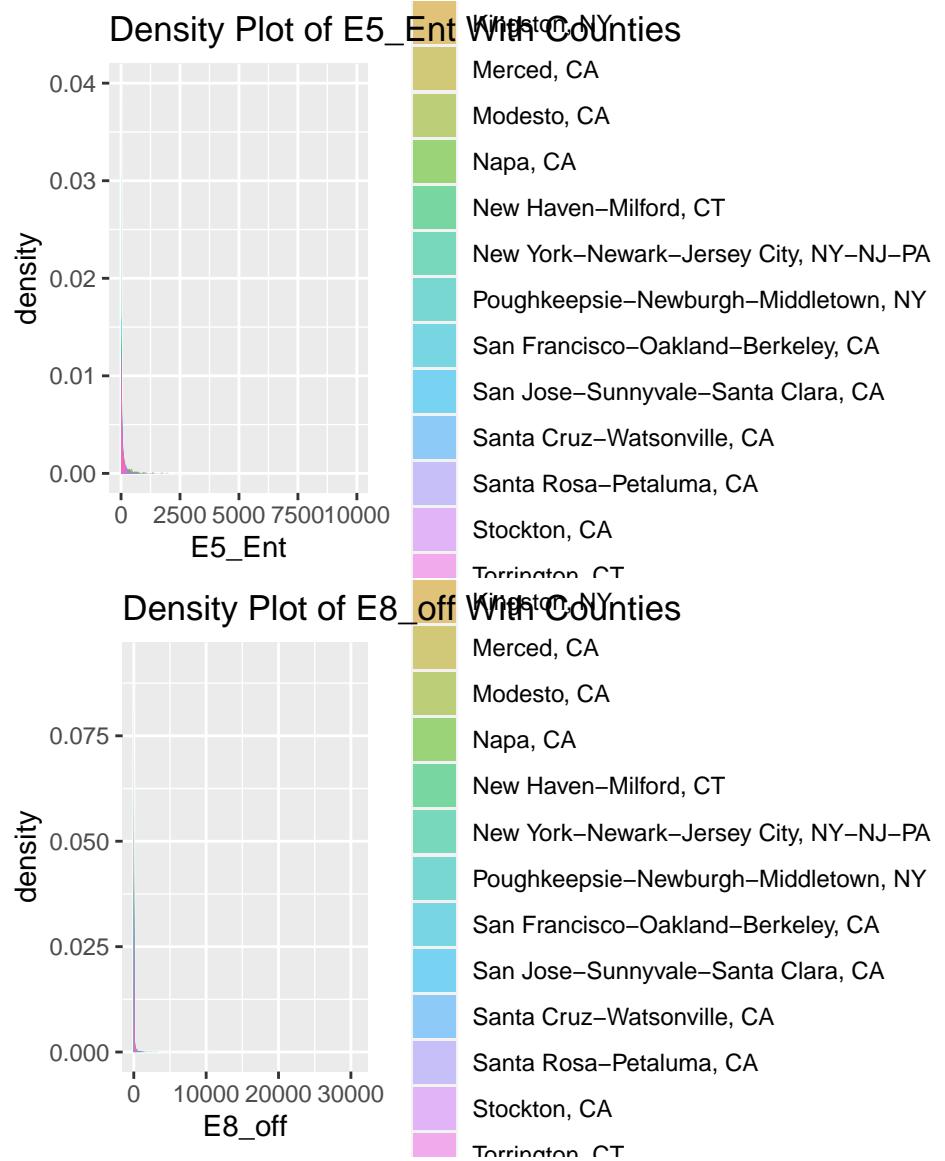


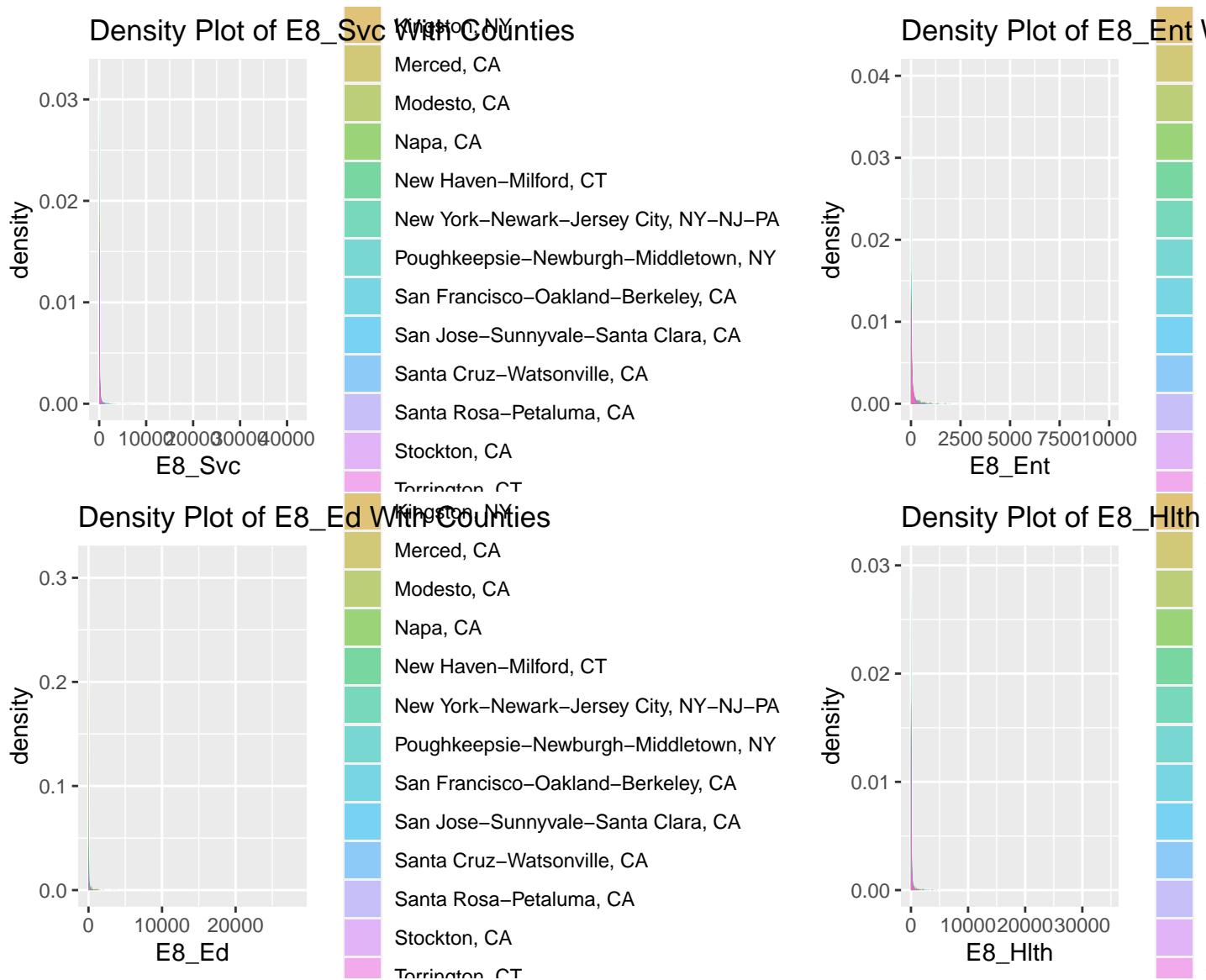


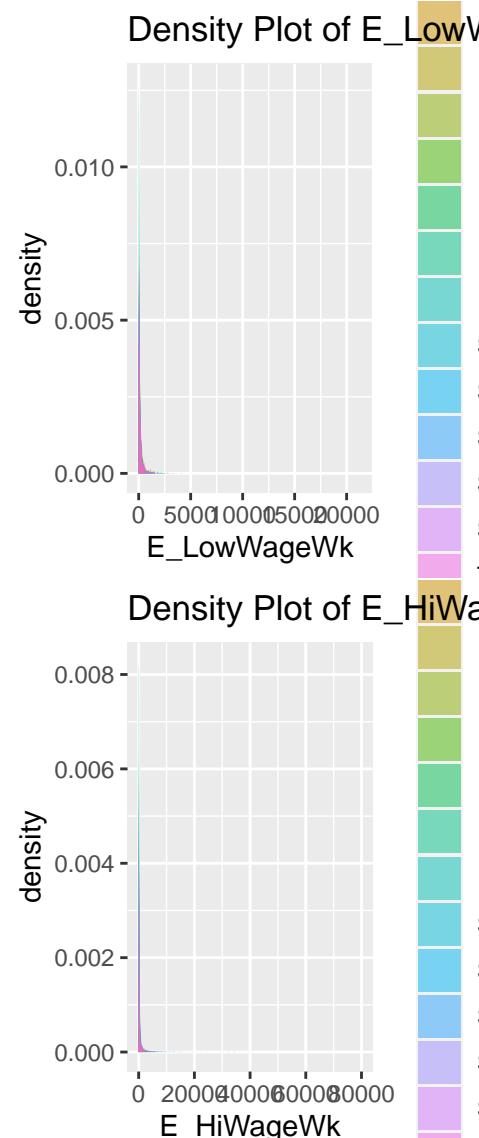
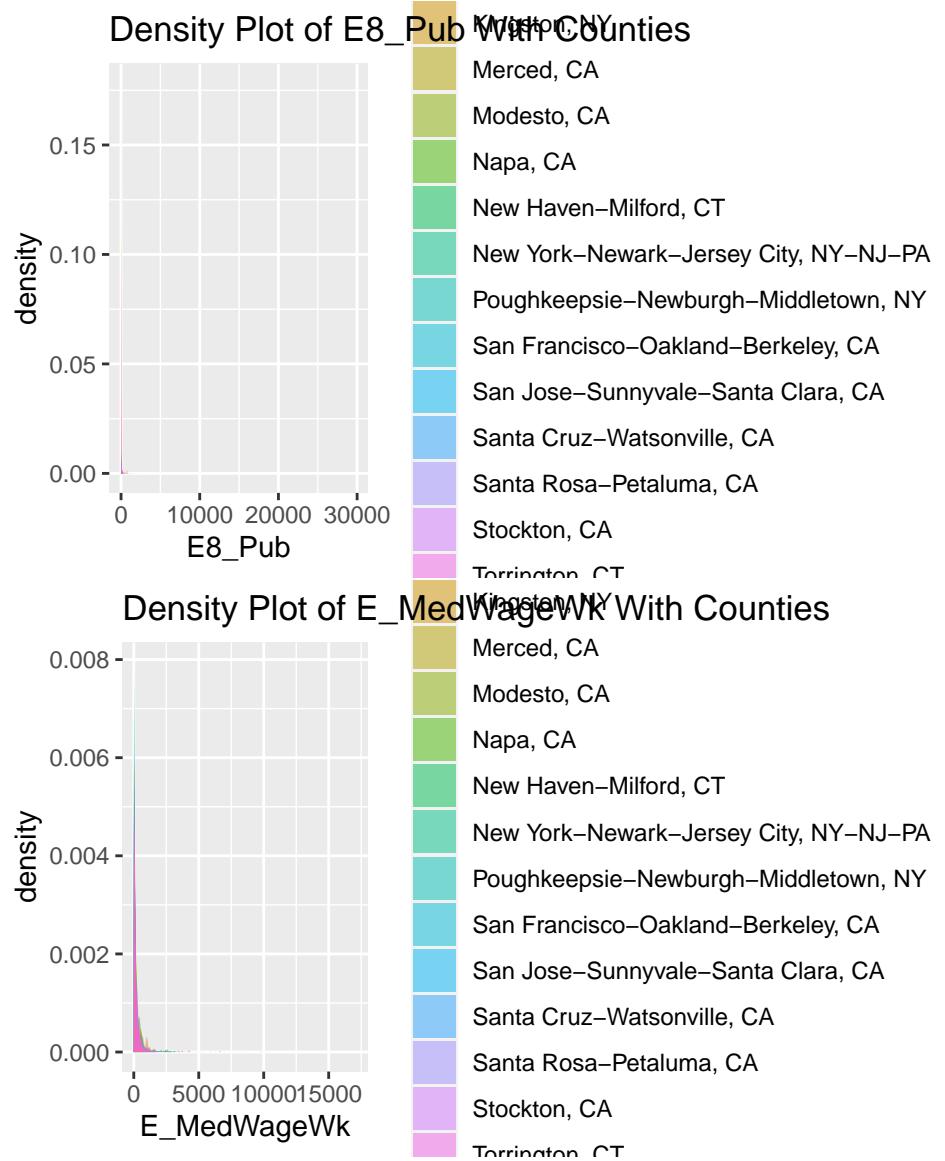


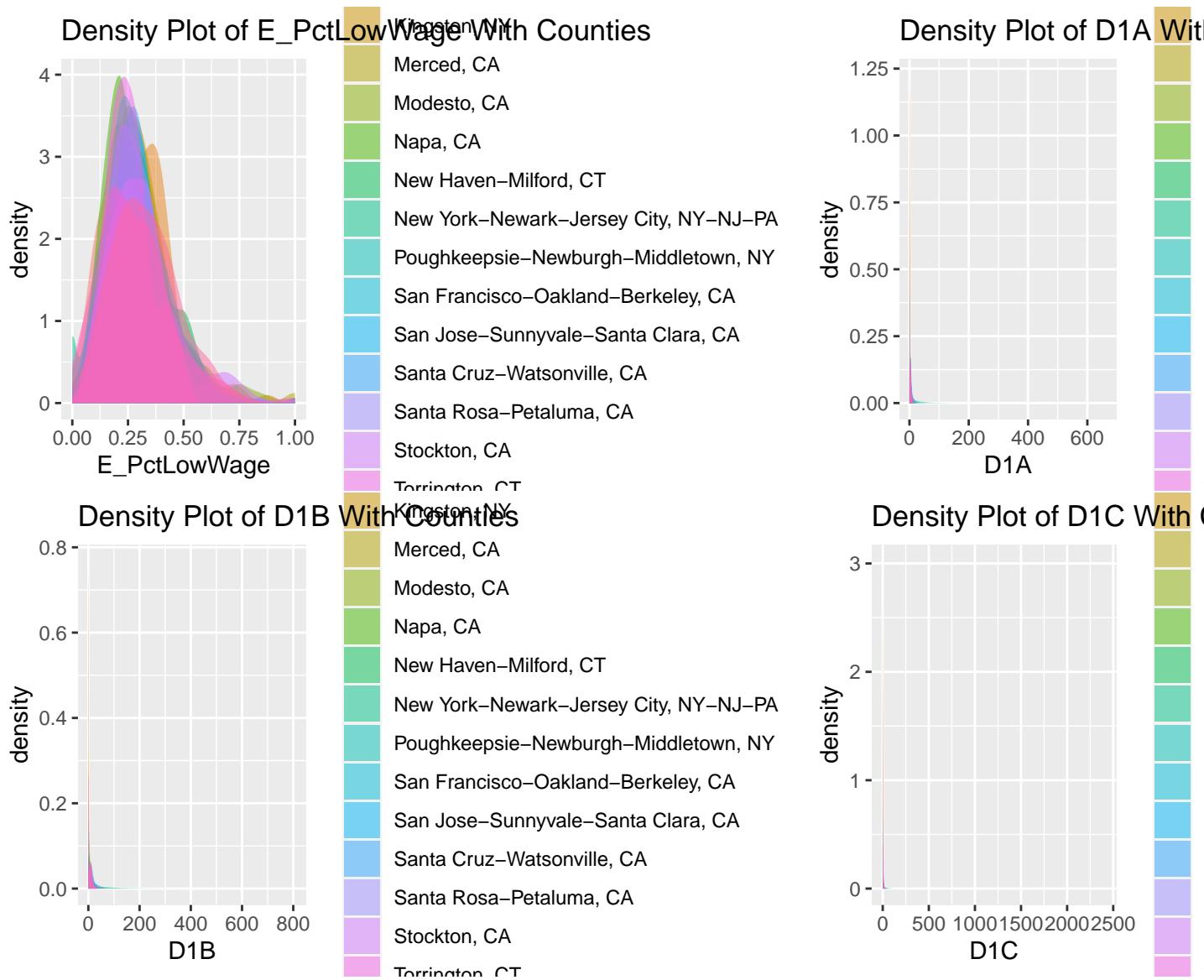
```
create_density_plots_all(SLD, 34, 54)
```



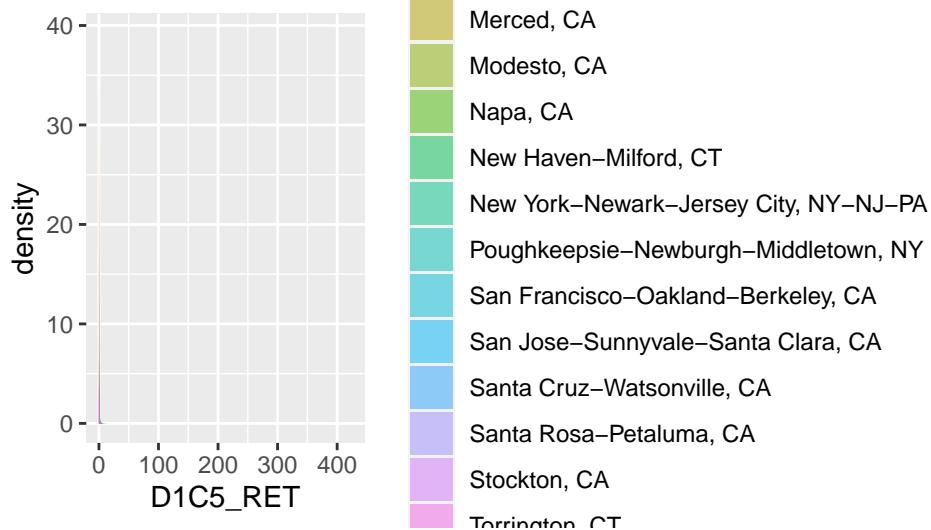






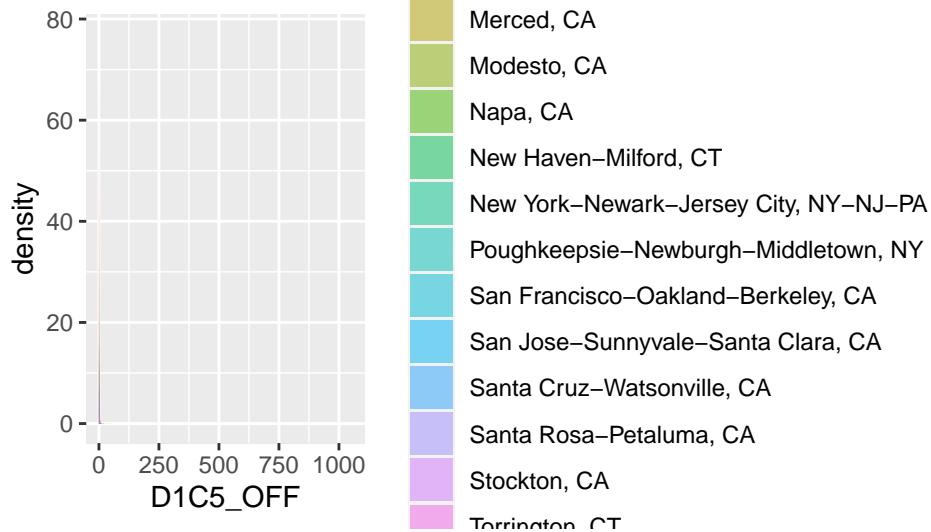


Density Plot of D1C5_RET With Counties

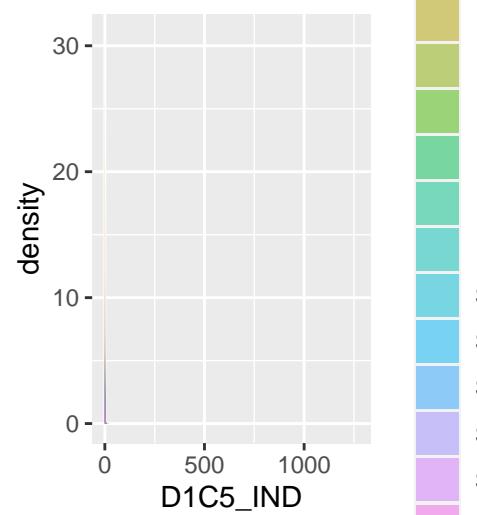


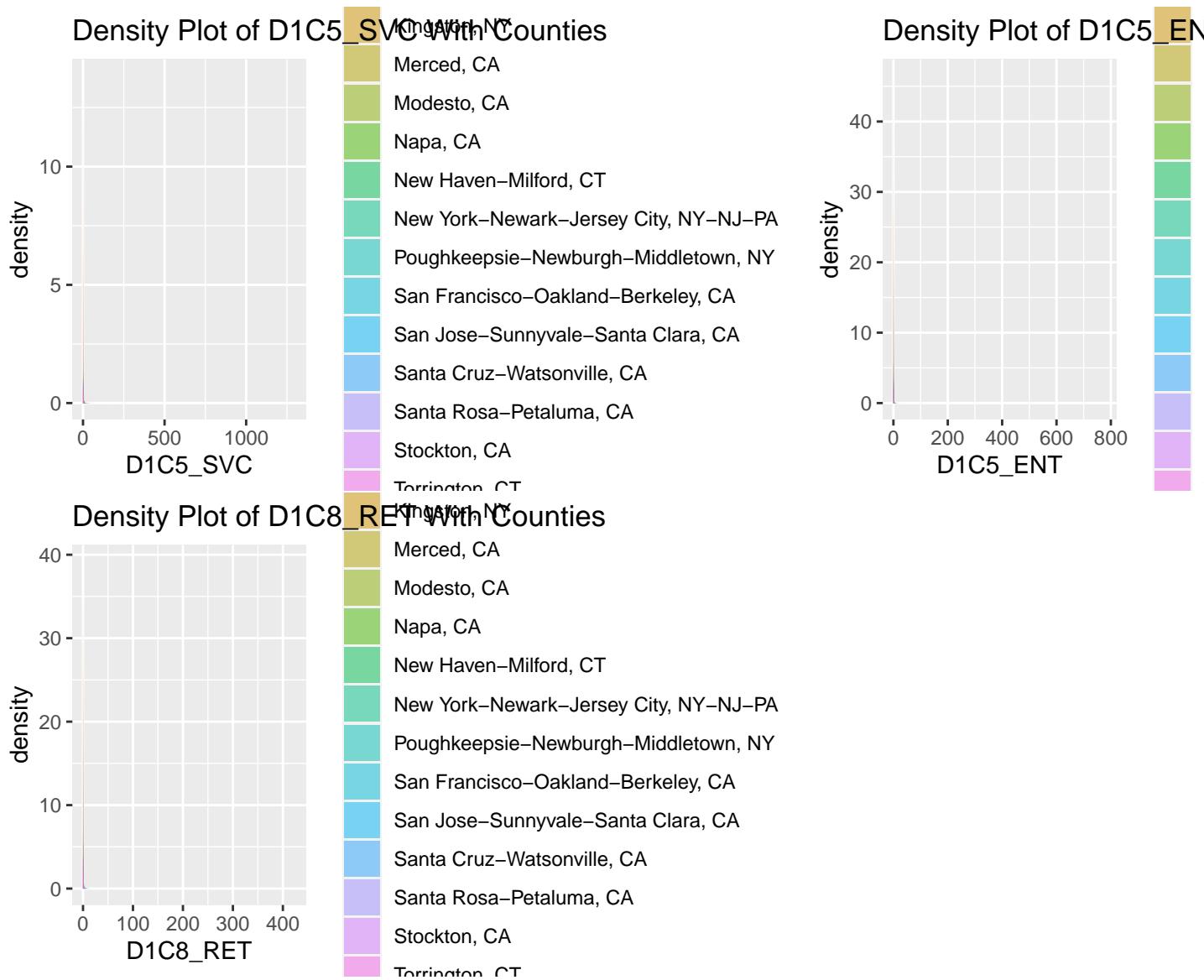
```
create_density_plots_all(SLD, 55, 76)
```

Density Plot of D1C5_OFF With Counties

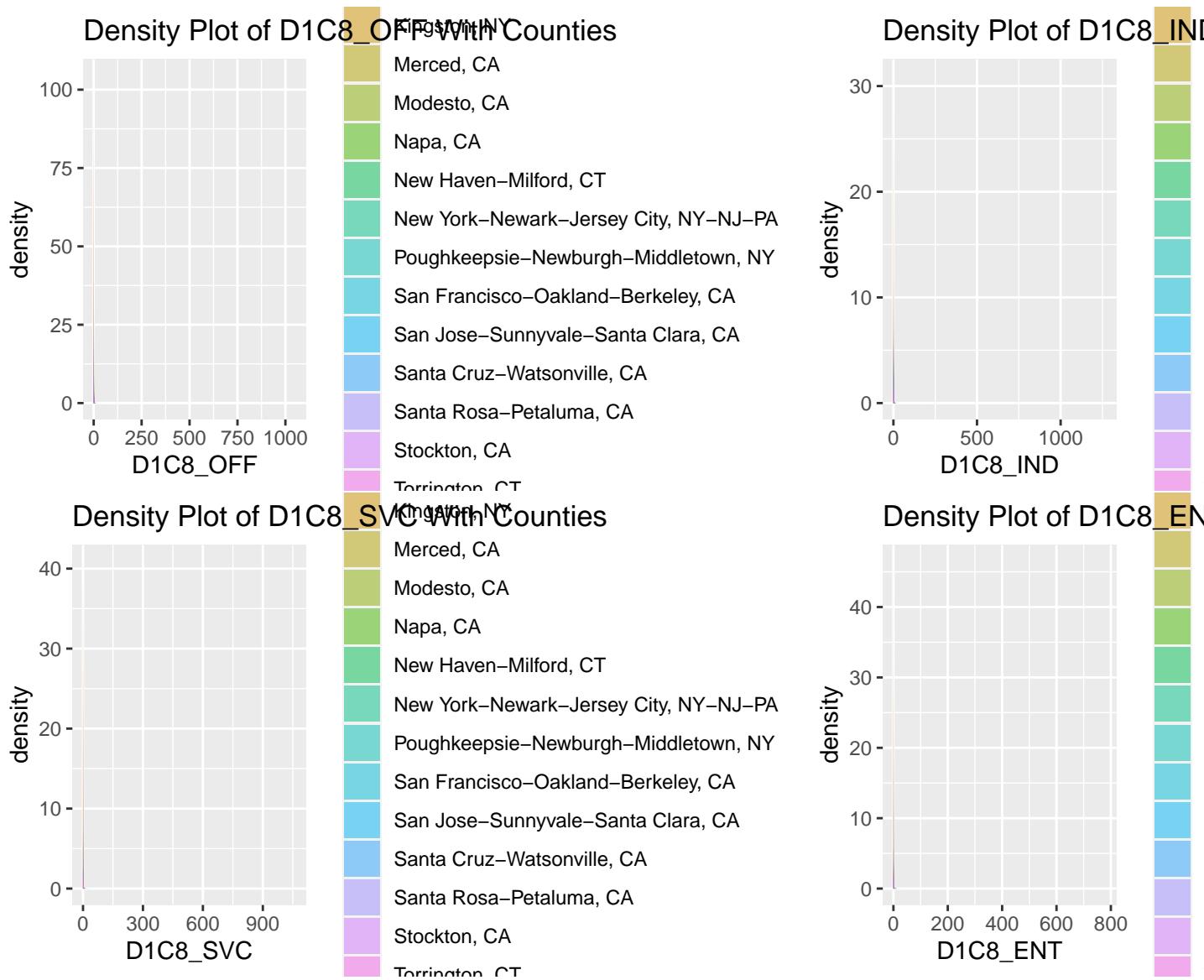


Density Plot of D1C5_IND With Counties

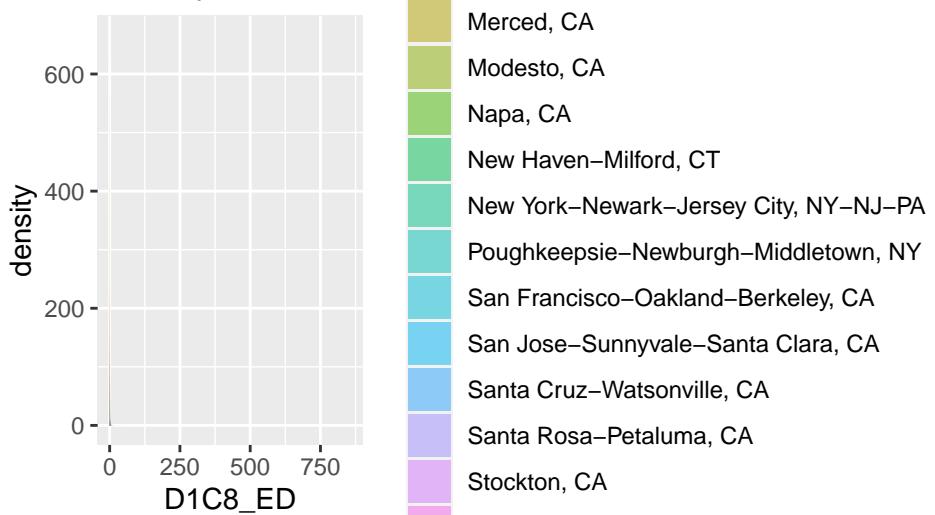




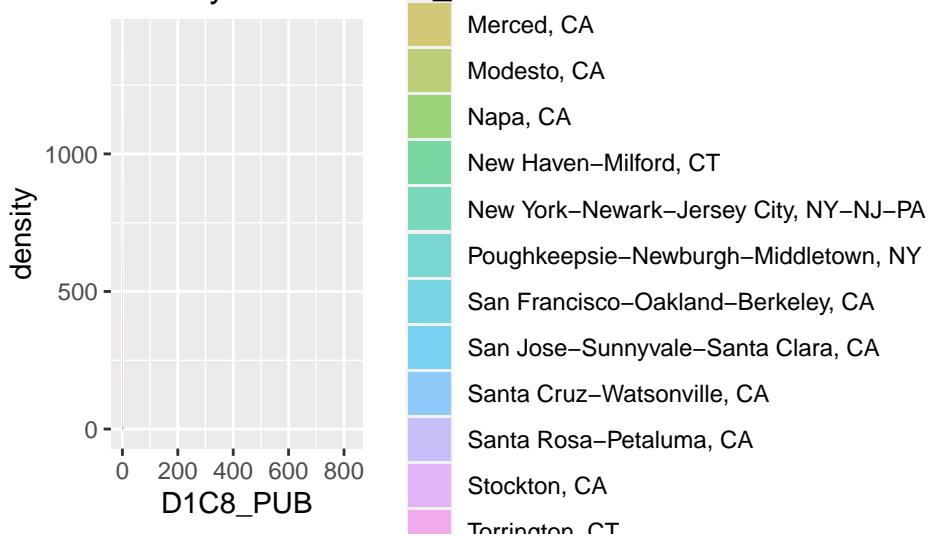
```
## Warning: Removed 232 rows containing non-finite values ('stat_density()').
```



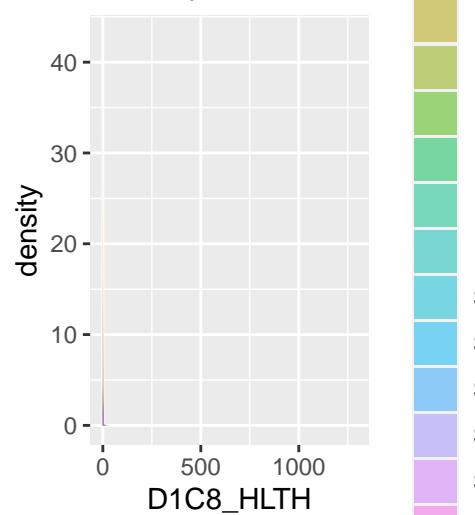
Density Plot of D1C8_ED With NY Counties



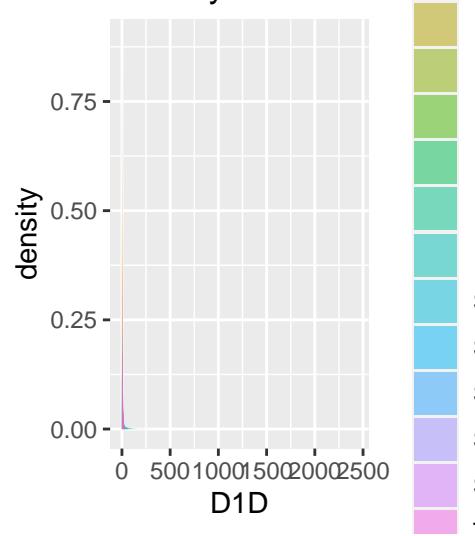
Density Plot of D1C8_PUB With NY Counties

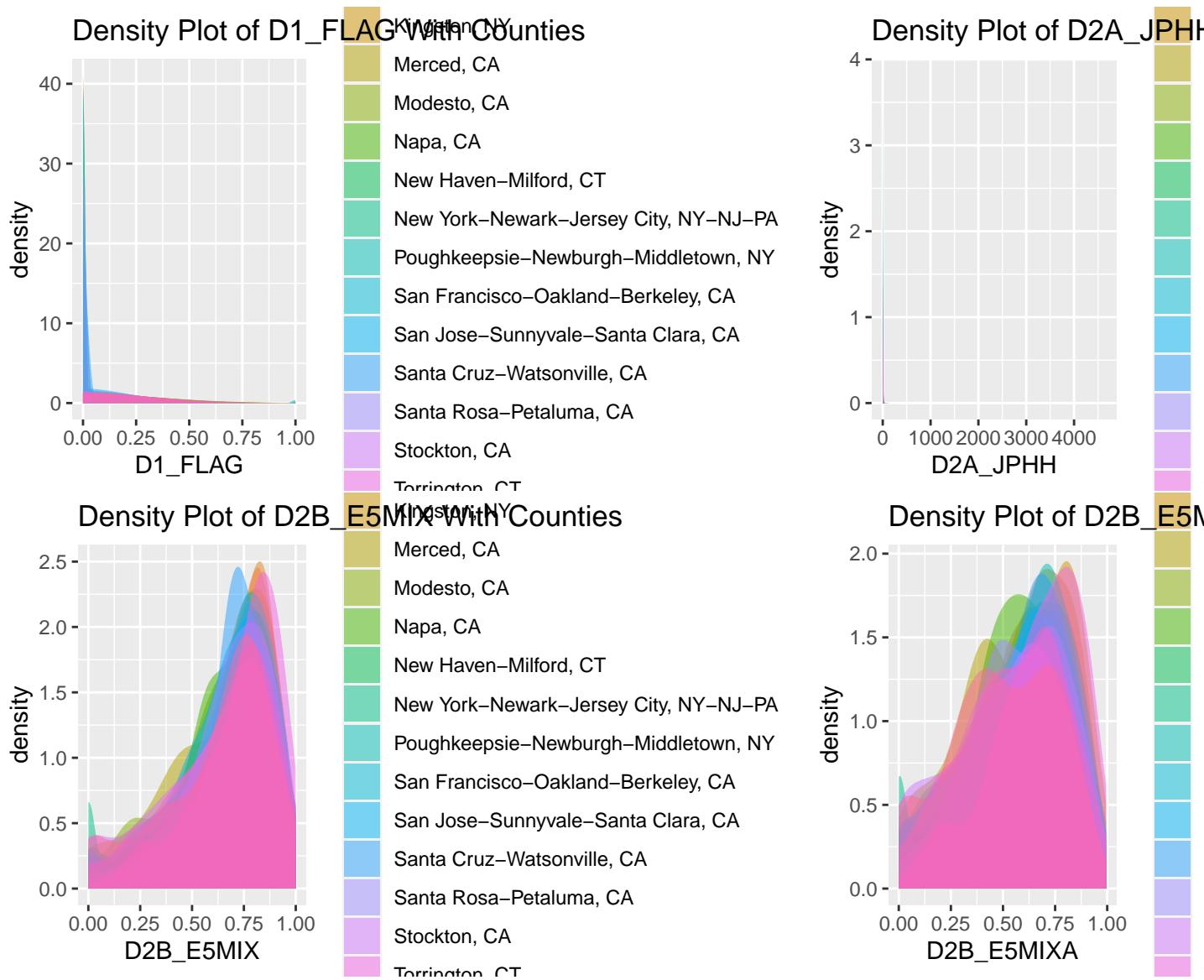


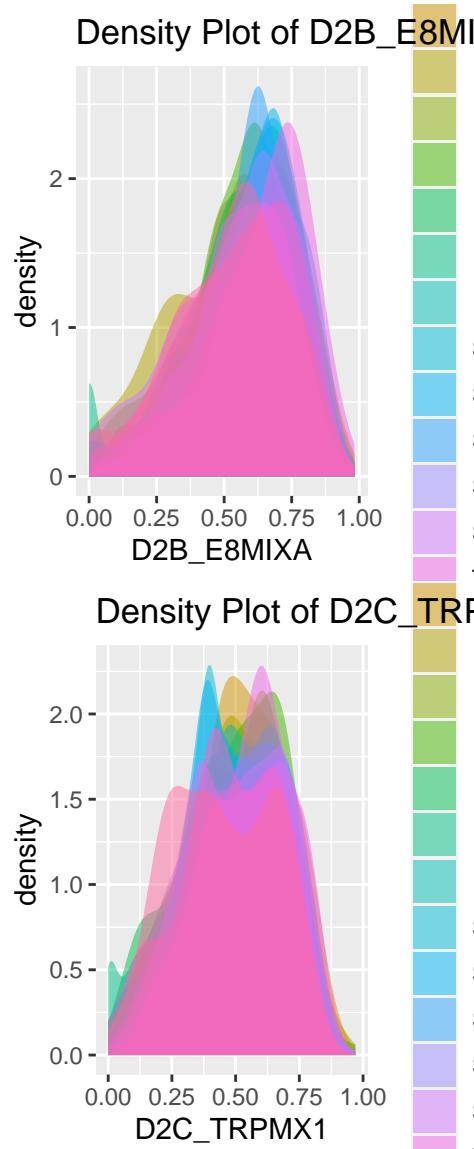
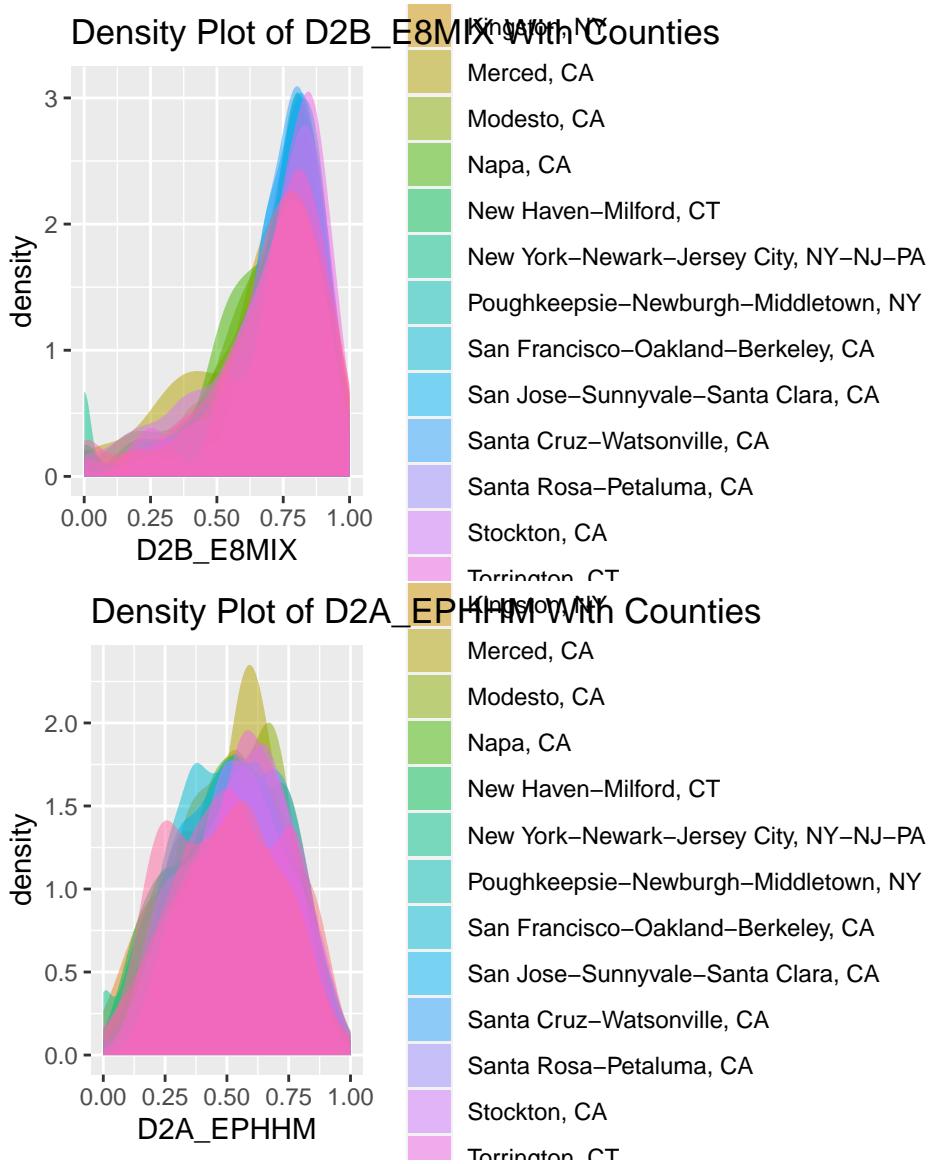
Density Plot of D1C8_HLTH With NY Counties



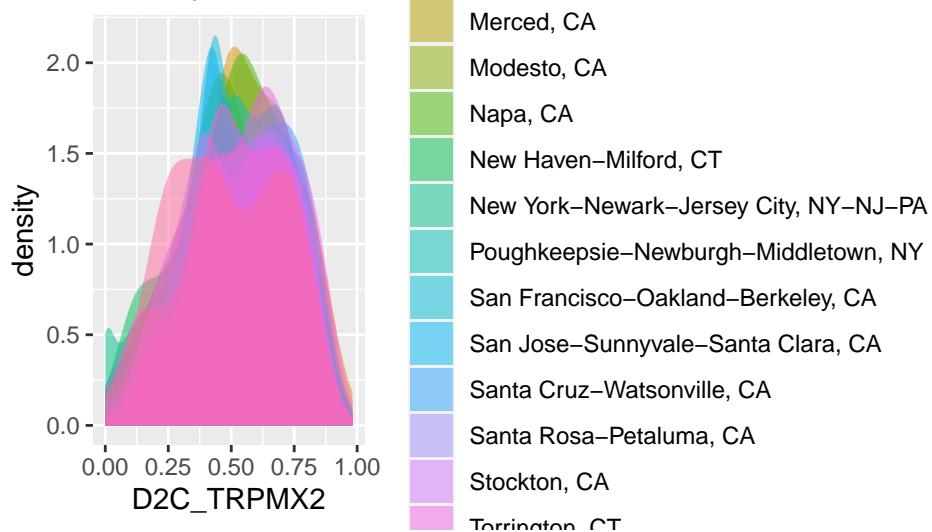
Density Plot of D1D With NY Counties





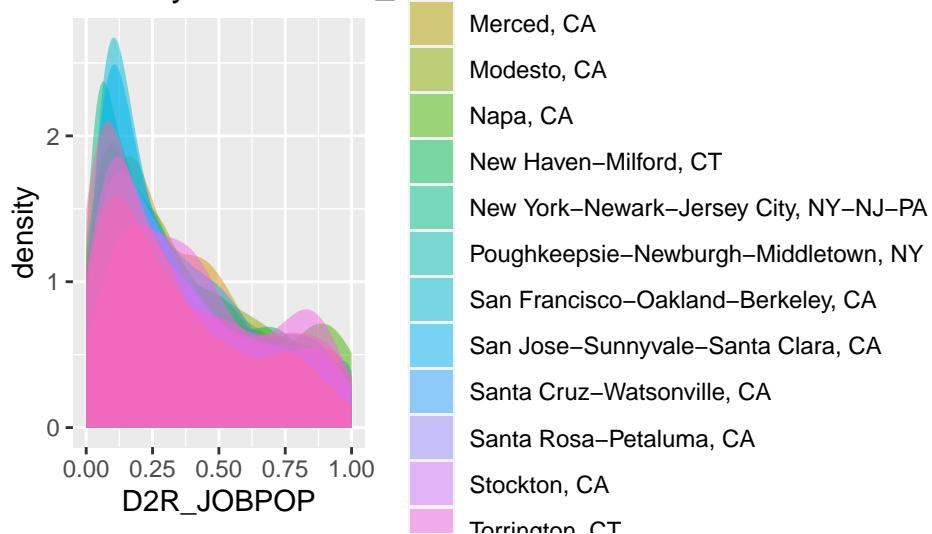


Density Plot of D2C_TRPMX2 With Counties

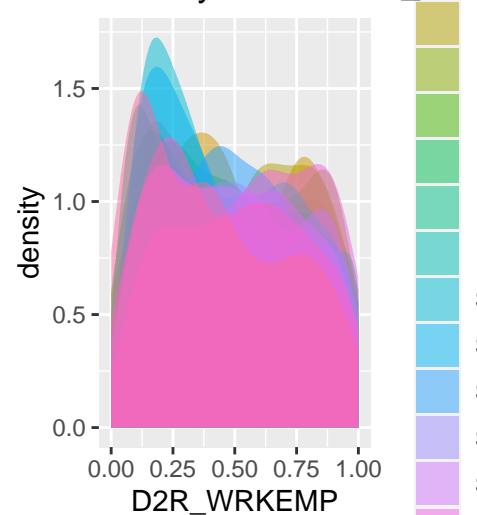


```
create_density_plots_all(SLD, 78, 98)
```

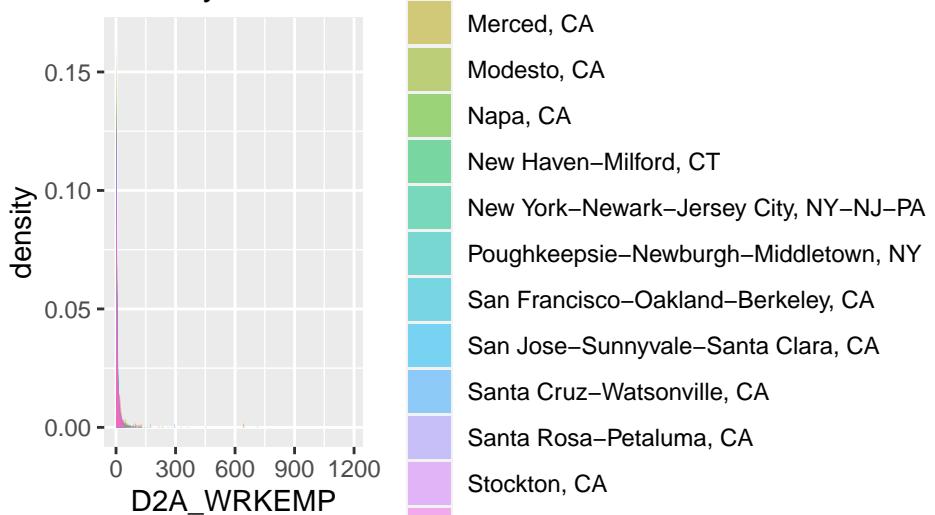
Density Plot of D2R_JOBPOP With Counties



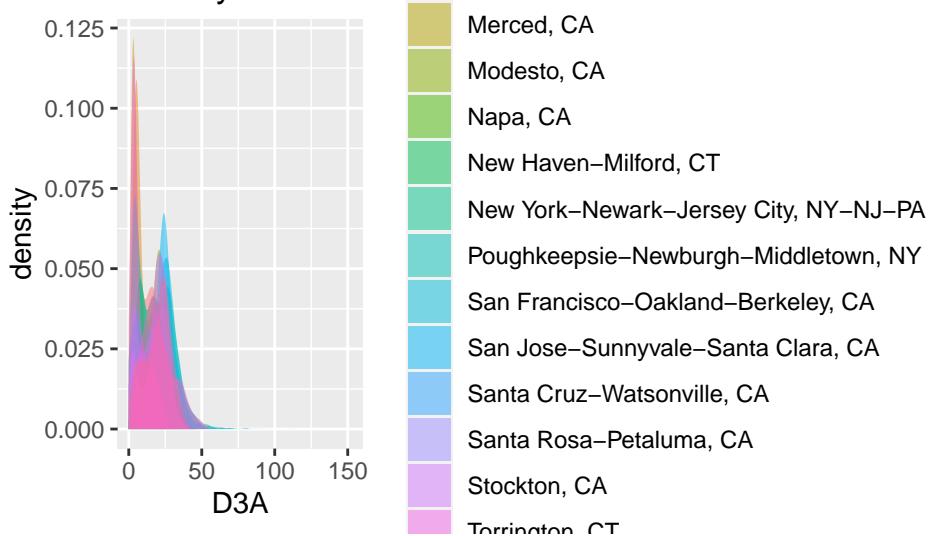
Density Plot of D2R_WRKEMP With Counties



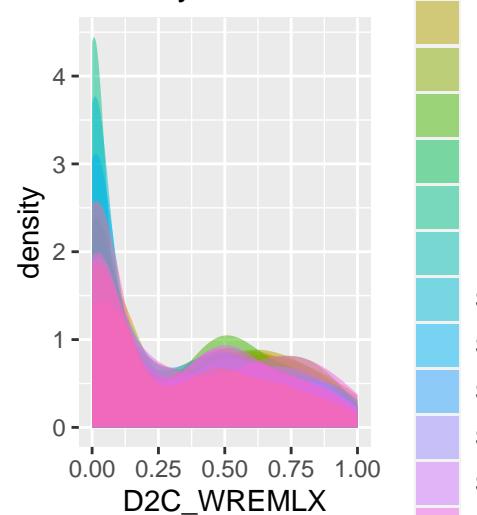
Density Plot of D2A_WRKEMP With Counties



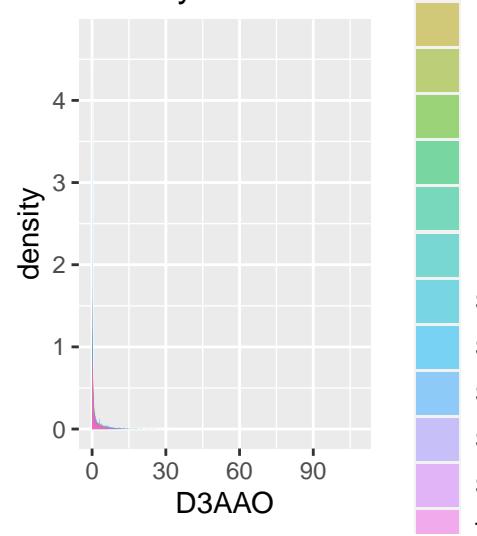
Density Plot of D3A With Counties

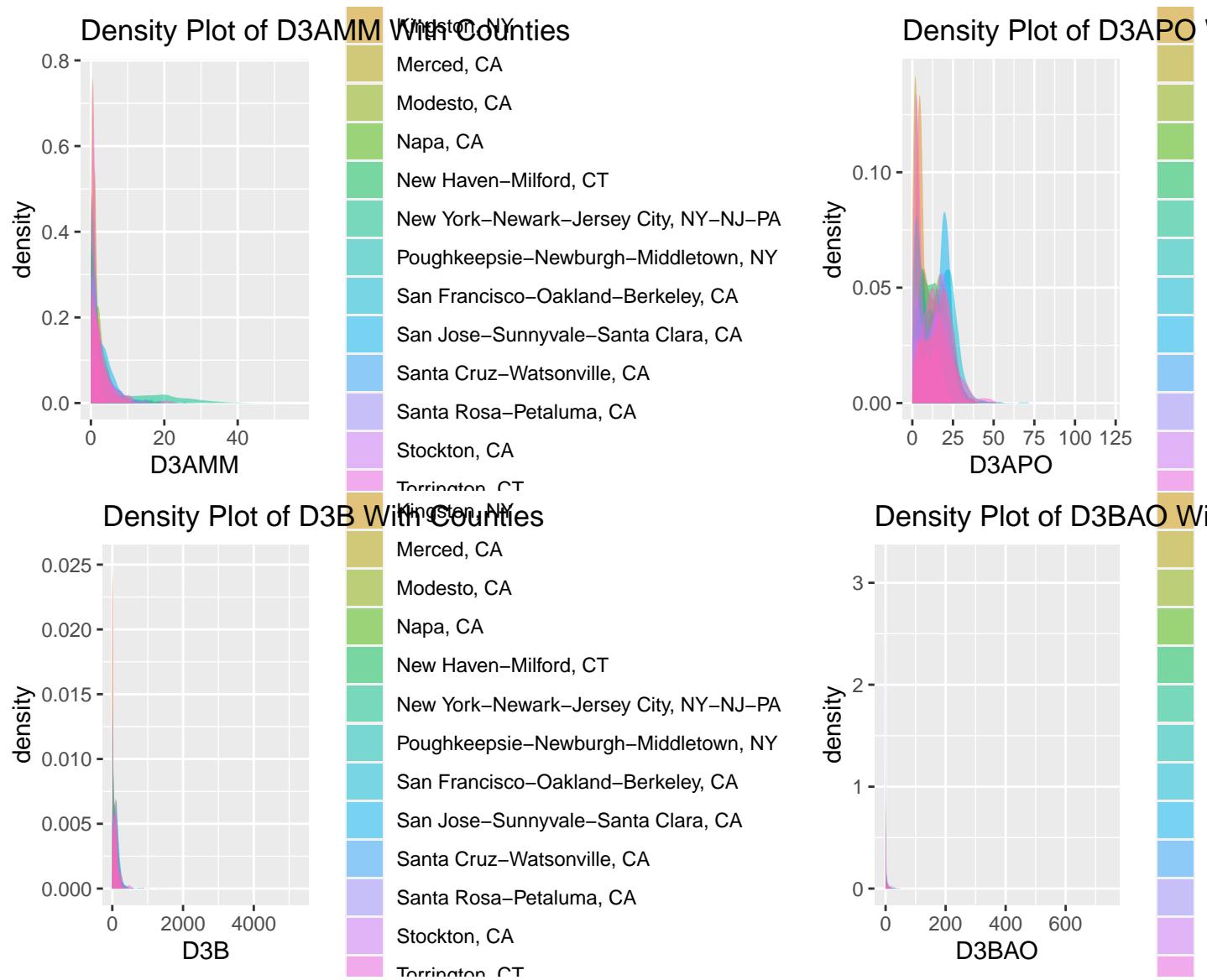


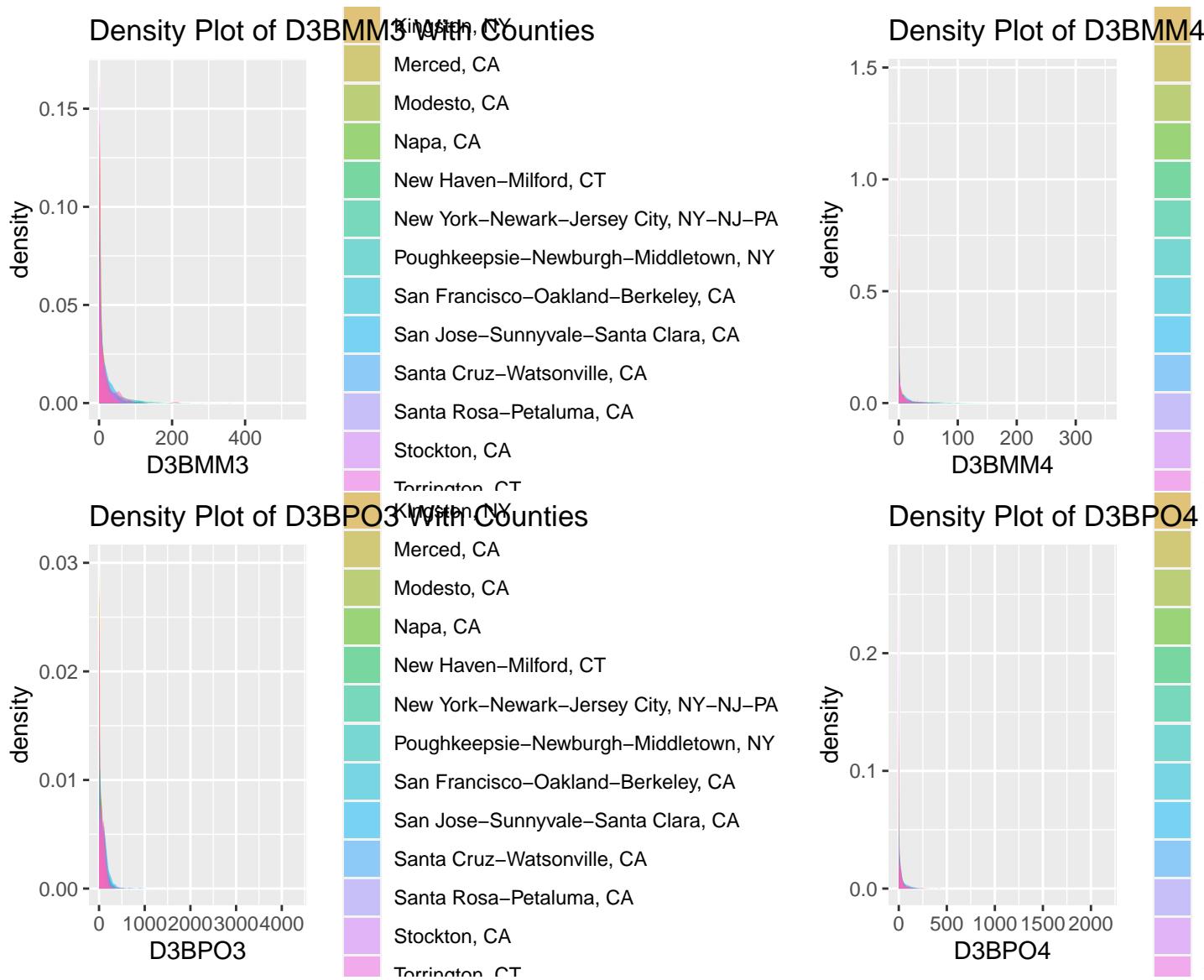
Density Plot of D2C_WREMLX

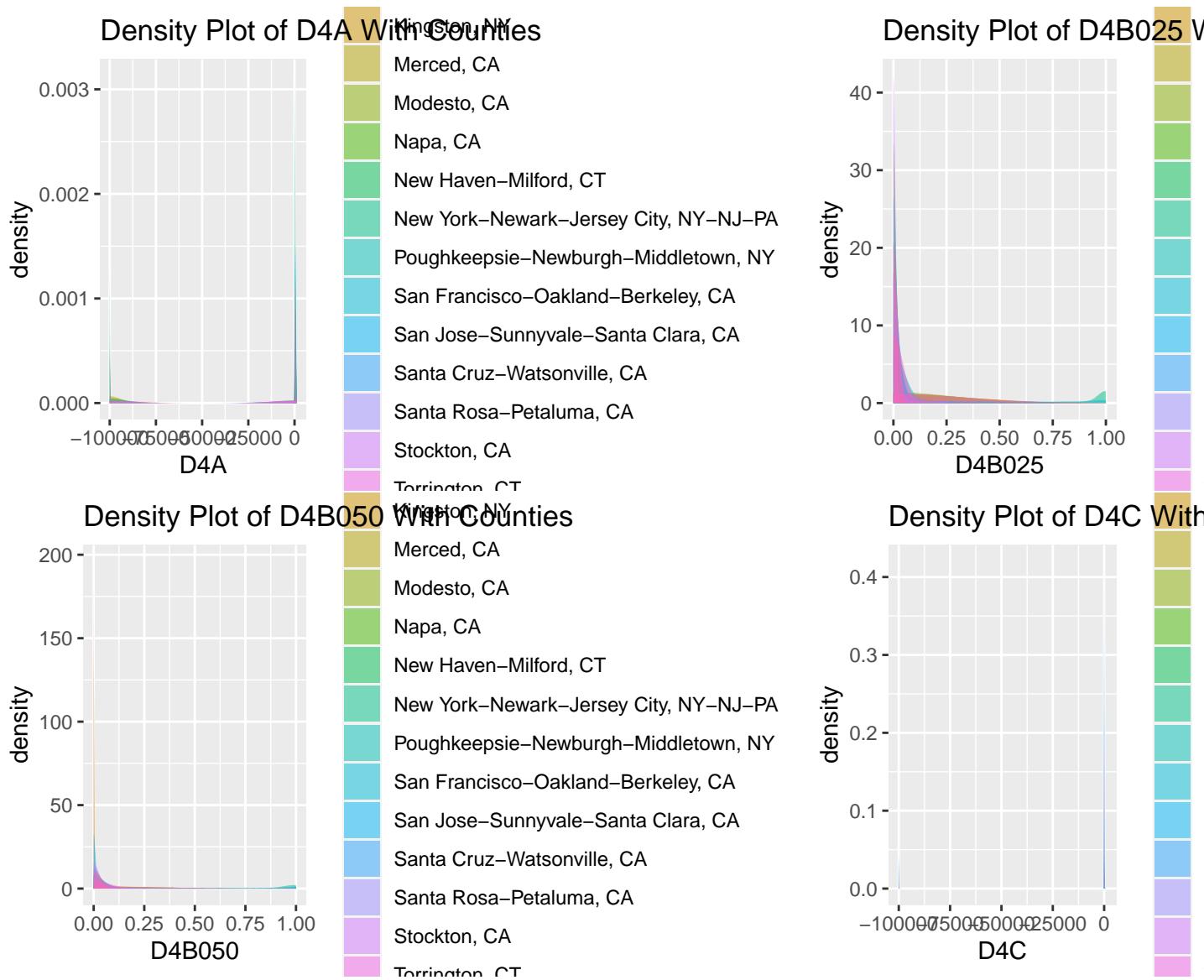


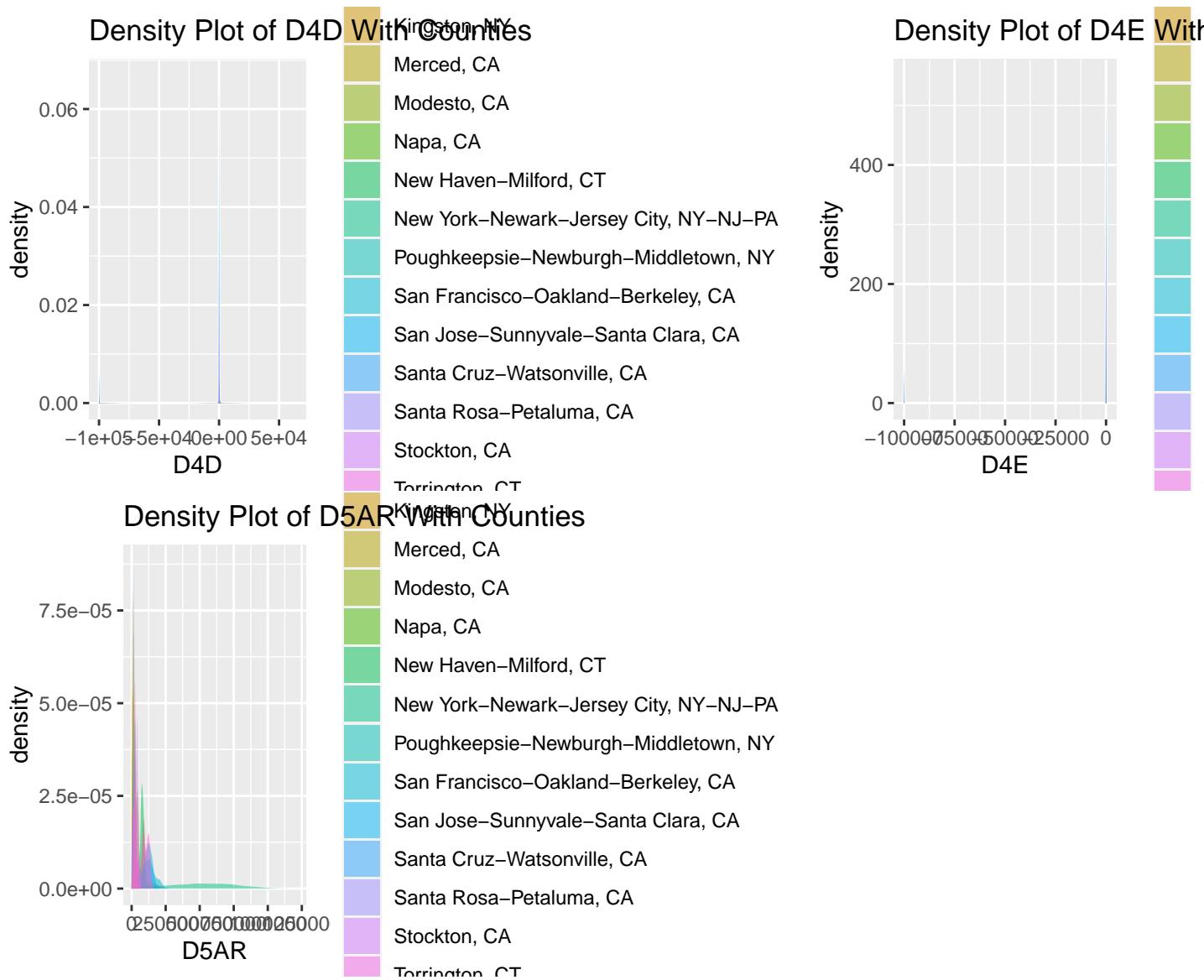
Density Plot of D3AAO With Counties





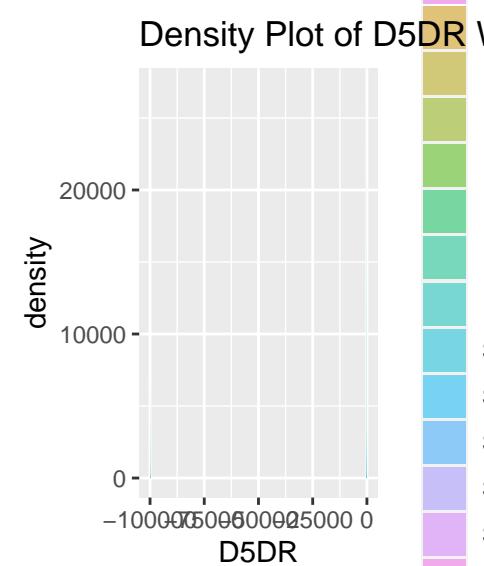
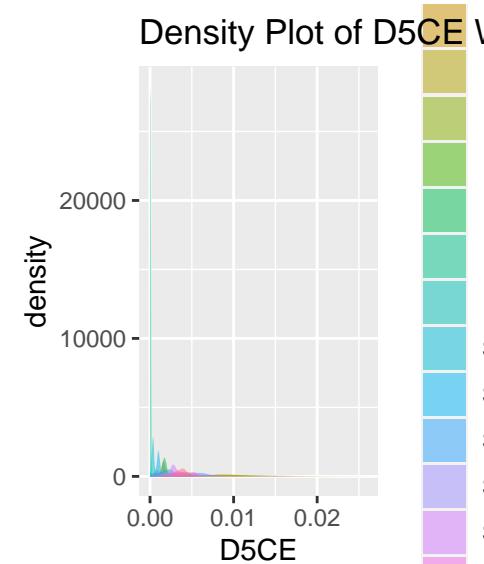
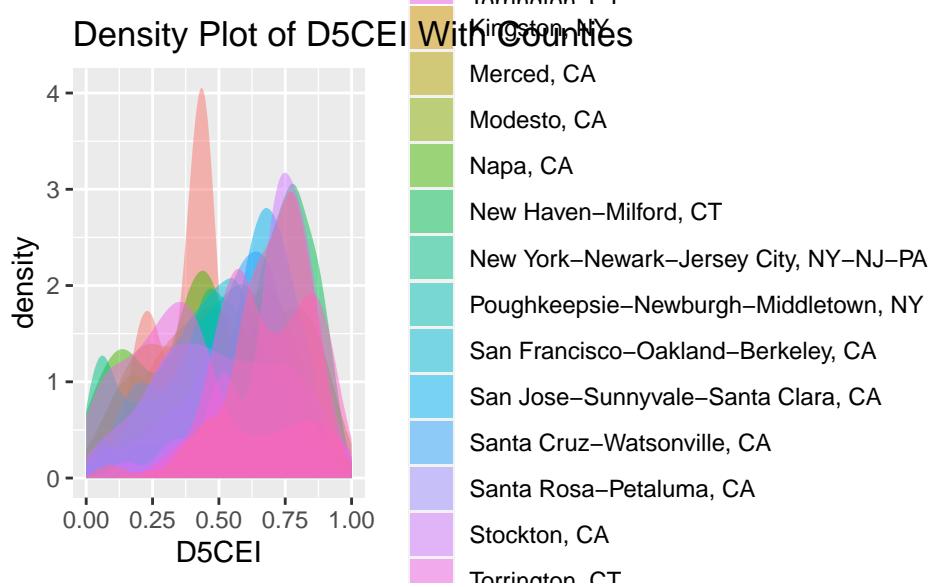
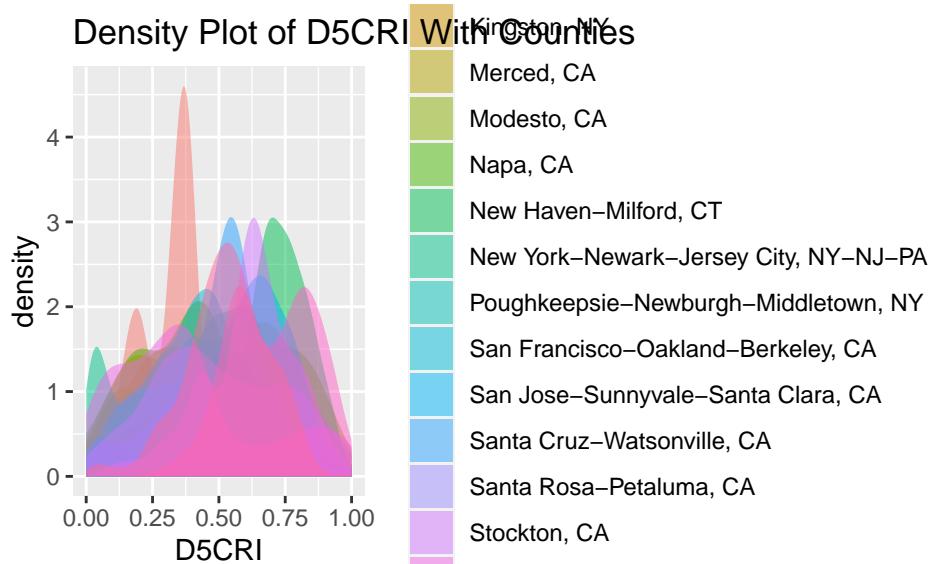




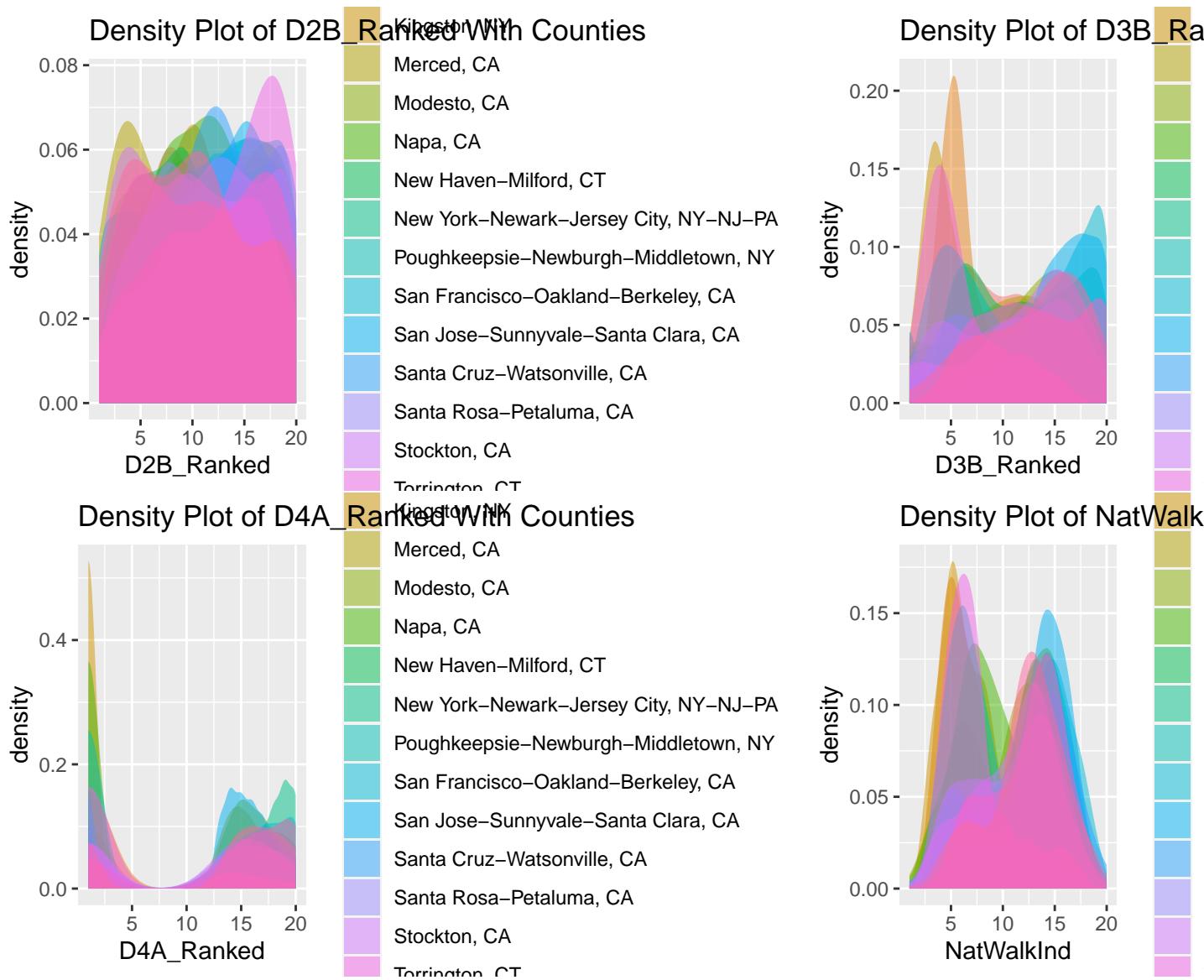


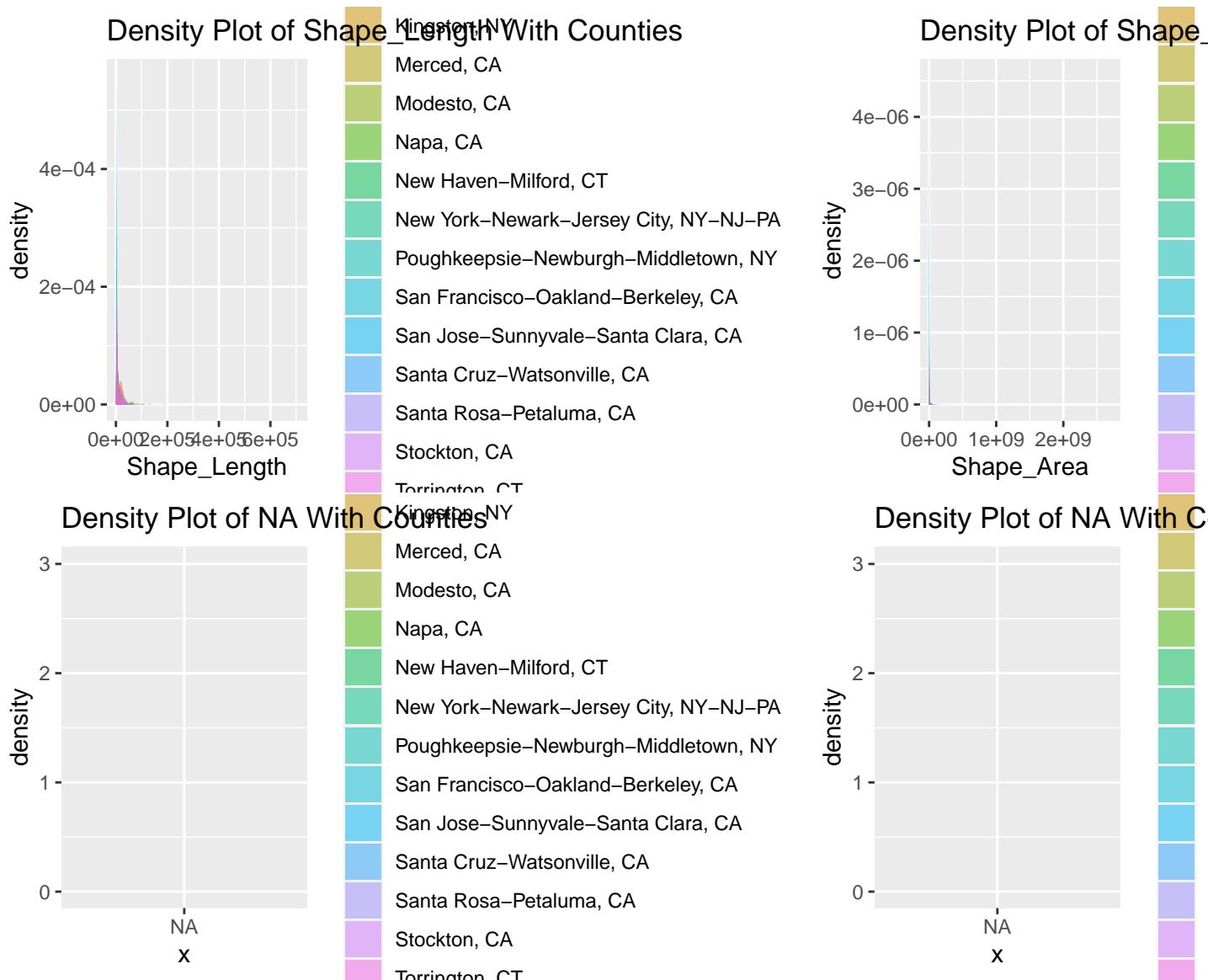
```
create_density_plots_all(SLD, 99, 118)
```











```
unique(SLD$CBSA_Name)
```

```
## [1] "Santa Rosa-Petaluma, CA"
## [2] "San Francisco-Oakland-Berkeley, CA"
## [3] "San Jose-Sunnyvale-Santa Clara, CA"
## [4] "Vallejo, CA"
## [5] "Napa, CA"
## [6] "Modesto, CA"
## [7] "Merced, CA"
## [8] "Stockton, CA"
## [9] "Santa Cruz-Watsonville, CA"
## [10] "New Haven-Milford, CT"
## [11] "Bridgeport-Stamford-Norwalk, CT"
## [12] "Torrington, CT"
## [13] "New York-Newark-Jersey City, NY-NJ-PA"
## [14] "Trenton-Princeton, NJ"
```

```

## [15] "Poughkeepsie-Newburgh-Middletown, NY"
## [16] "Kingston, NY"
## [17] "East Stroudsburg, PA"

unique(SLD$CSA_Name)

## [1] "San Jose-San Francisco-Oakland, CA" "New York-Newark, NY-NJ-CT-PA"

NewYork_SanFrancisco_data <- read_csv("data/NewYork_SanFrancisco_data.csv", show_col_types = FALSE)

## New names:
## * ' ' -> '...1'

unique(NewYork_SanFrancisco_data$CBSA_Name)

## [1] "Santa Rosa-Petaluma, CA"
## [2] "San Francisco-Oakland-Berkeley, CA"
## [3] "San Jose-Sunnyvale-Santa Clara, CA"
## [4] "Vallejo, CA"
## [5] "Napa, CA"
## [6] "Modesto, CA"
## [7] "Merced, CA"
## [8] "Stockton, CA"
## [9] "Santa Cruz-Watsonville, CA"
## [10] "New Haven-Milford, CT"
## [11] "Bridgeport-Stamford-Norwalk, CT"
## [12] "Torrington, CT"
## [13] "New York-Newark-Jersey City, NY-NJ-PA"
## [14] "Trenton-Princeton, NJ"
## [15] "Poughkeepsie-Newburgh-Middletown, NY"
## [16] "Kingston, NY"
## [17] "East Stroudsburg, PA"

unique(NewYork_SanFrancisco_data$CSA_Name)

## [1] "San Jose-San Francisco-Oakland, CA" "New York-Newark, NY-NJ-CT-PA"

names(NewYork_SanFrancisco_data)

## [1] "...1"          "OBJECTID"      "GEOID10"        "GEOID20"        "STATEFP"
## [6] "COUNTYFP"      "TRACTCE"        "BLKGRPCE"       "CSA"            "CSA_Name"
## [11] "CBSA"          "CBSA_Name"       "CBSA_POP"        "CBSA_EMP"       "CBSA_WRK"
## [16] "Ac_Total"      "Ac_Water"       "Ac_Land"         "Ac_Unpr"        "TotPop"
## [21] "CountHU"       "HH"              "P_WrkAge"        "AutoOwn0"       "Pct_A00"
## [26] "AutoOwn1"      "Pct_A01"         "AutoOwn2p"       "Pct_A02p"        "Workers"
## [31] "R_LowWageWk"   "R_MedWageWk"   "R_HiWageWk"     "R_PCTLOWWAGE"  "TotEmp"
## [36] "E5_Ret"         "E5_Off"          "E5_Ind"          "E5_Svc"          "E5_Ent"
## [41] "E8_Ret"         "E8_off"          "E8_Ind"          "E8_Svc"          "E8_Ent"
## [46] "E8_Ed"          "E8_Hlth"         "E8_Pub"          "E_LowWageWk"    "E_MedWageWk"
## [51] "E_HiWageWk"    "E_PctLowWage"  "D1A"             "D1B"            "D1C"

```

```

## [56] "D1C5_RET"      "D1C5_OFF"       "D1C5_IND"       "D1C5_SVC"       "D1C5_ENT"
## [61] "D1C8_RET"      "D1C8_OFF"       "D1C8_IND"       "D1C8_SVC"       "D1C8_ENT"
## [66] "D1C8_ED"       "D1C8_HLTH"      "D1C8_PUB"       "D1D"           "D1_FLAG"
## [71] "D2A_JPHH"      "D2B_E5MIX"      "D2B_E5MIXA"     "D2B_E8MIX"      "D2B_E8MIXA"
## [76] "D2A_EPHHM"     "D2C_TRPMX1"     "D2C_TRPMX2"     "D2C_TRIPEQ"    "D2R_JOBPOP"
## [81] "D2R_WRKEMP"    "D2A_WRKEMP"     "D2C_WREMLX"     "D3A"           "D3AAO"
## [86] "D3AMM"          "D3APO"          "D3B"            "D3BAO"         "D3BMM3"
## [91] "D3BMM4"          "D3BP03"          "D3BP04"         "D4A"           "D4B025"
## [96] "D4B050"          "D4C"             "D4D"            "D4E"           "D5AR"
## [101] "D5AE"           "D5BR"           "D5BE"          "D5CR"          "D5CRI"
## [106] "D5CE"           "D5CEI"          "D5DR"          "D5DRI"         "D5DE"
## [111] "D5DEI"          "D2A_Ranked"     "D2B_Ranked"     "D3B_Ranked"     "D4A_Ranked"
## [116] "NatWalkInd"     "Shape_Length"    "Shape_Area"

```

```

#removing missing values.
NewYork_SanFrancisco_data <- na.omit(NewYork_SanFrancisco_data)

```

#Univariate Analysis:

```

create_density_plots_subset <- function(data, start_col, end_col) {
  numeric_vars <- names(data)[sapply(data, is.numeric)]
  selected_vars <- numeric_vars[start_col:end_col]

  for (var in selected_vars) {
    # Create density plot for each variable
    plot <- gf_density(as.formula(paste("~", var)), data = data, fill = ~CSA_Name,
                        title = paste("Density Plot of", var))
    print(plot)
  }
}

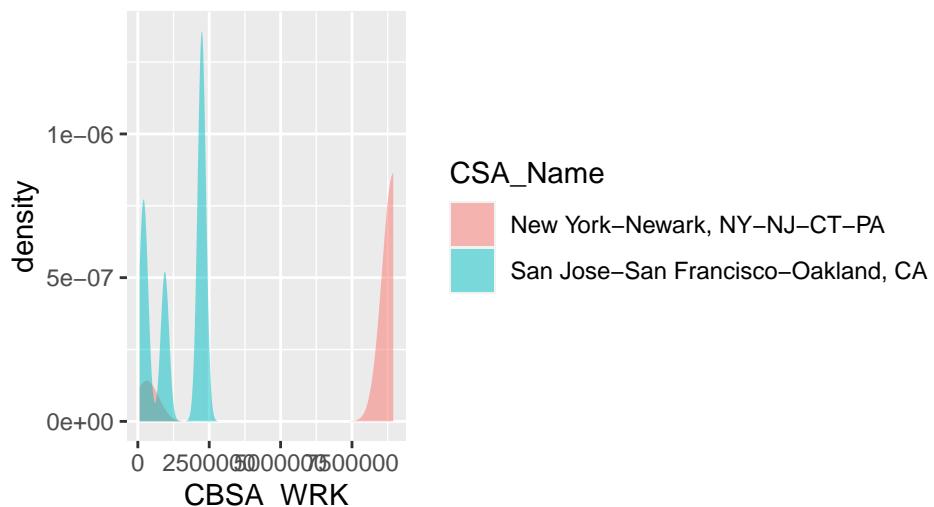
# Function to create density plots for all numeric variables

# Example usage:
# Assuming 'data' is your dataset and you want to create density plots for columns 1 to 10:

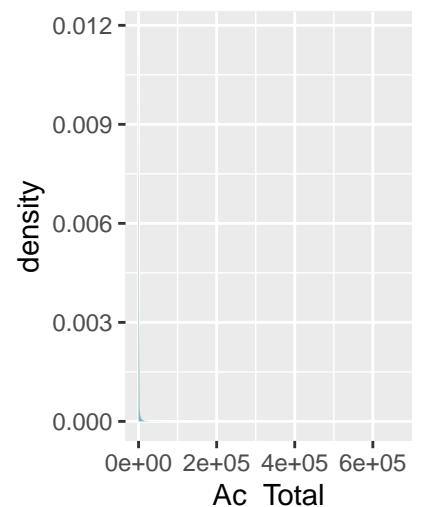
create_density_plots_subset(NewYork_SanFrancisco_data, 13, 33)

```

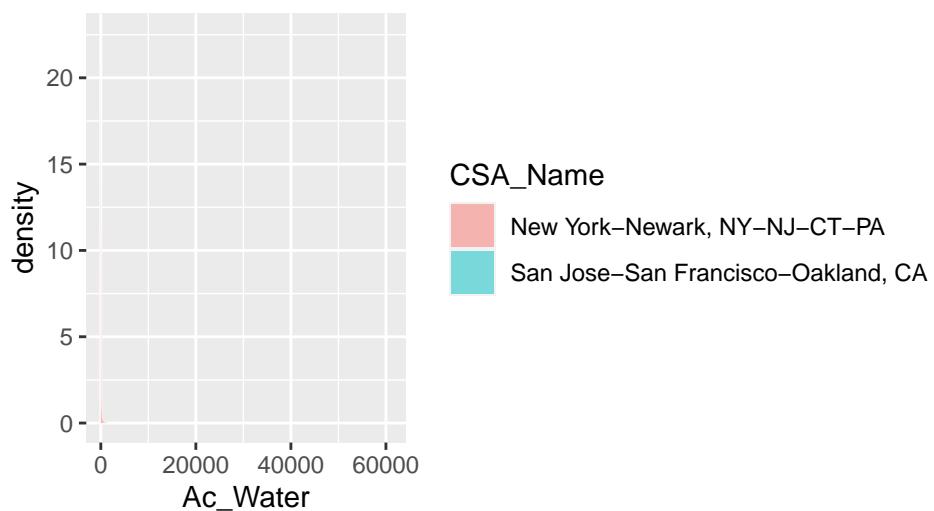
Density Plot of CBSA_WRK



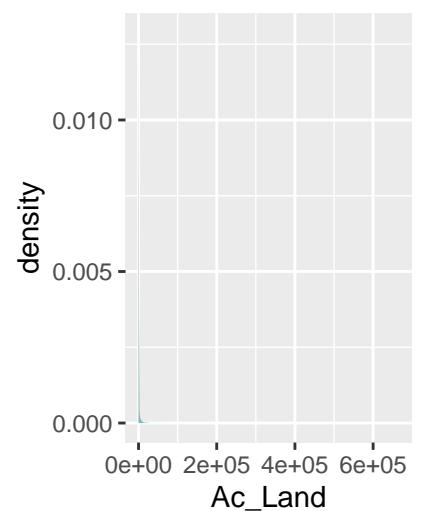
Density Plot of Ac_Total



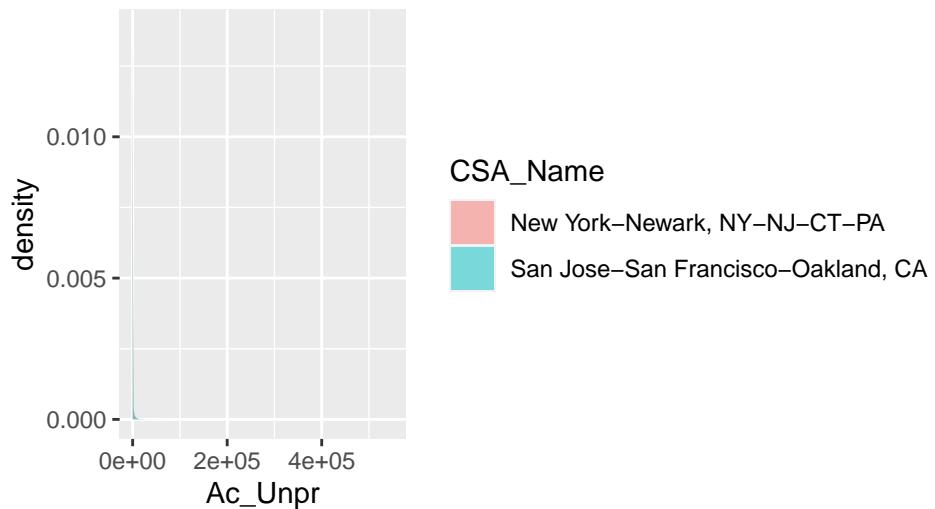
Density Plot of Ac_Water



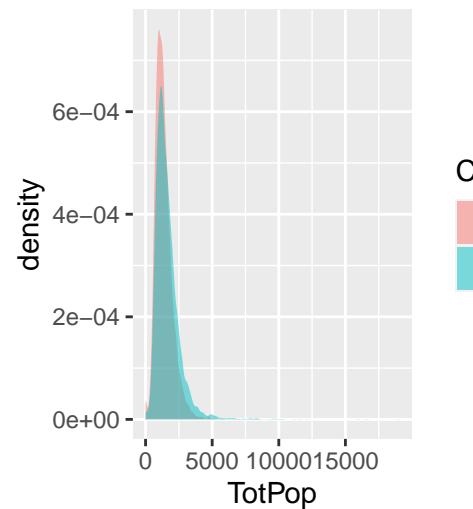
Density Plot of Ac_Land



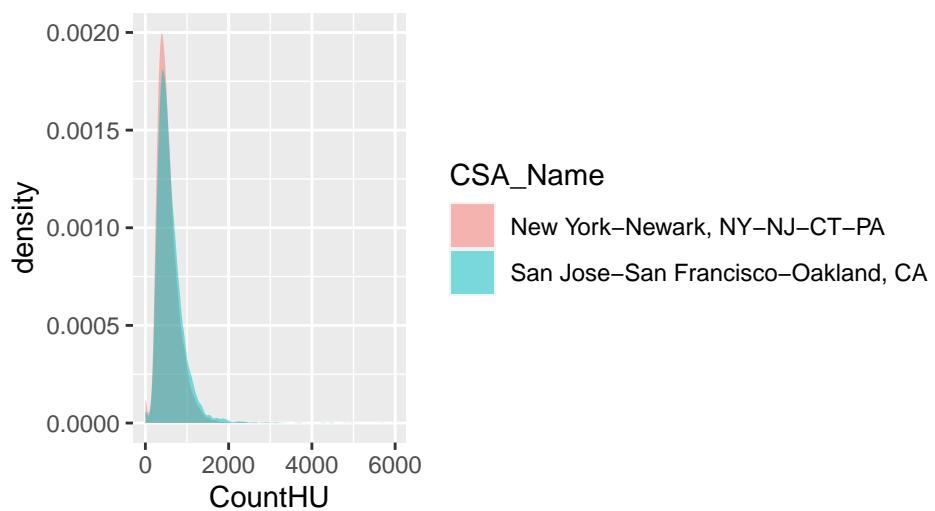
Density Plot of Ac_Unpr



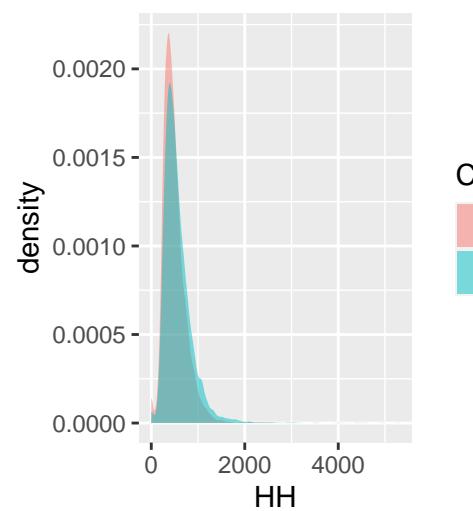
Density Plot of TotPop



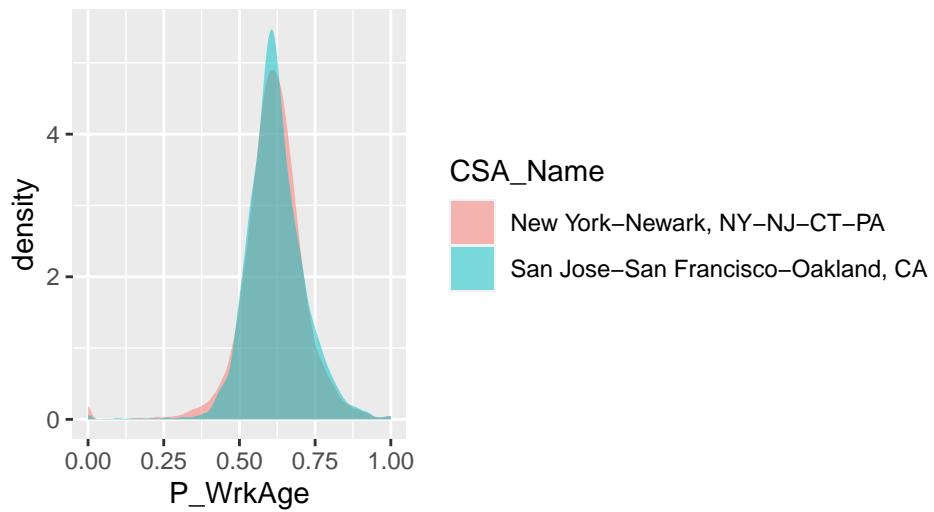
Density Plot of CountHU



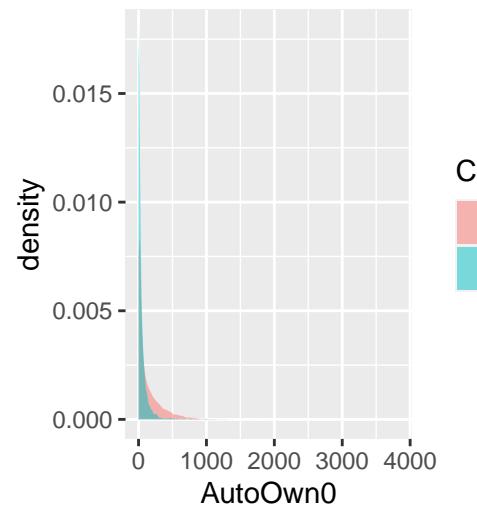
Density Plot of HH



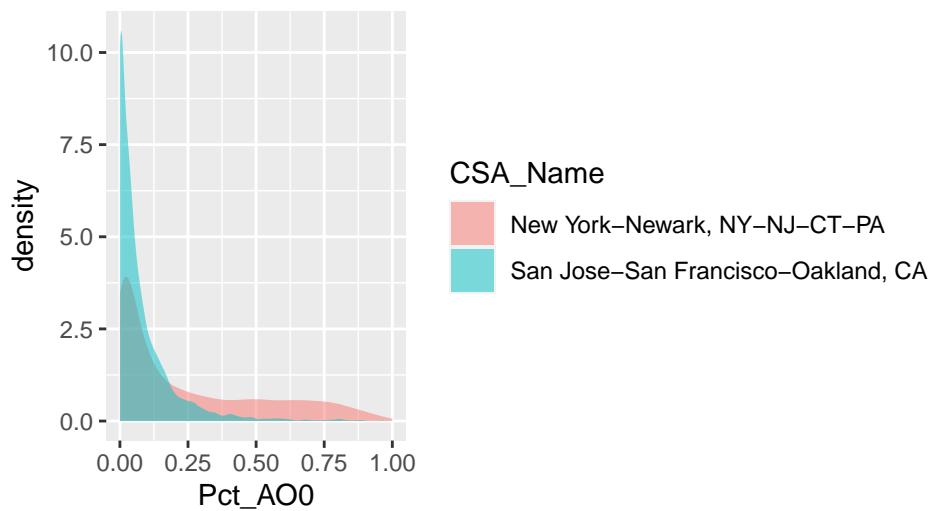
Density Plot of P_WrkAge



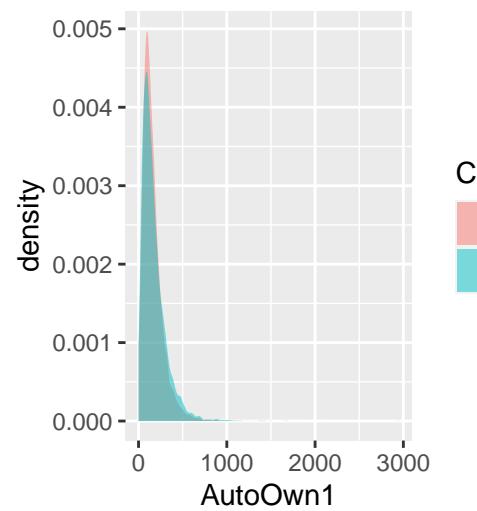
Density Plot of AutoOwn0



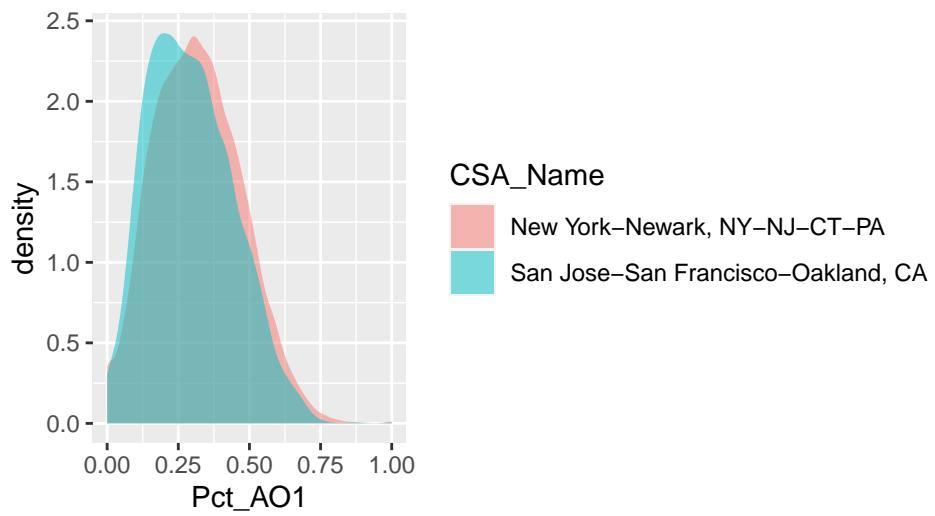
Density Plot of Pct_AO0



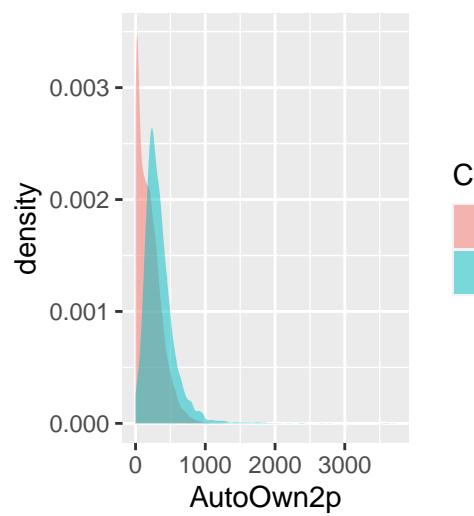
Density Plot of AutoOwn1



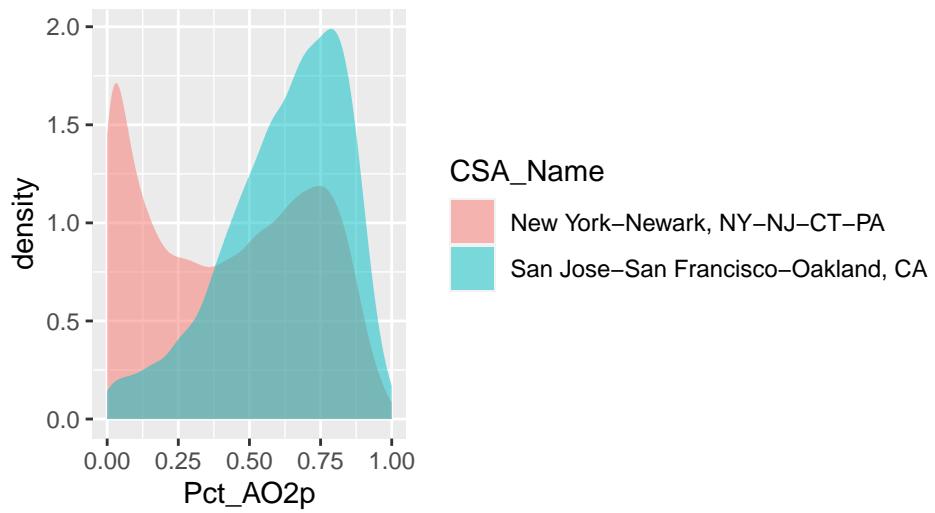
Density Plot of Pct_AO1



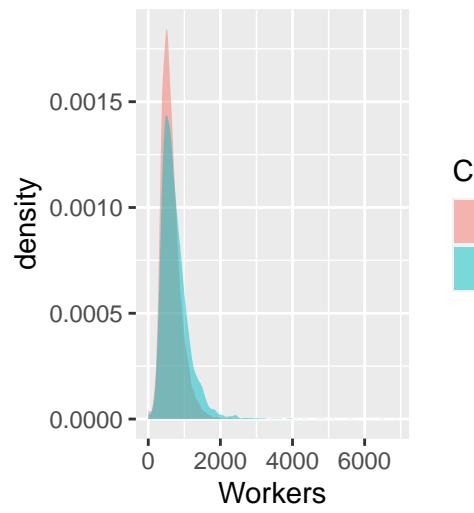
Density Plot of AutoOwn2p



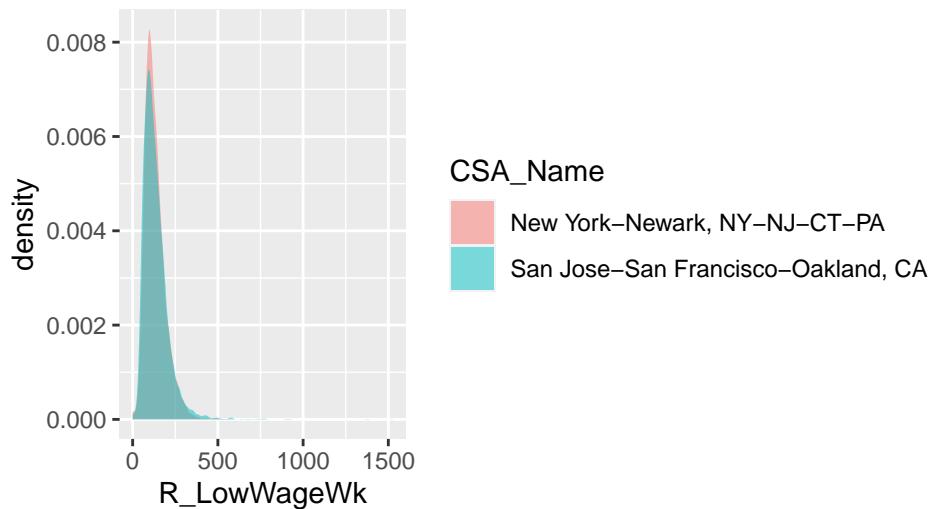
Density Plot of Pct_AO2p



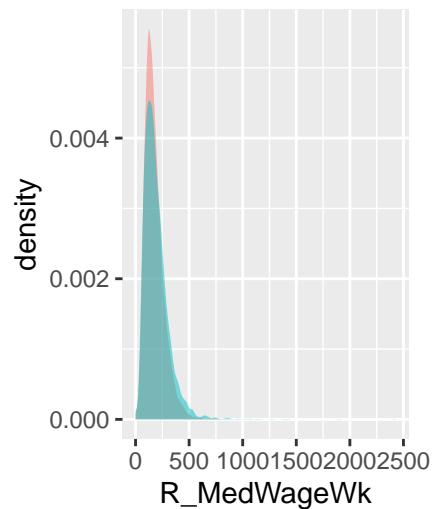
Density Plot of Workers



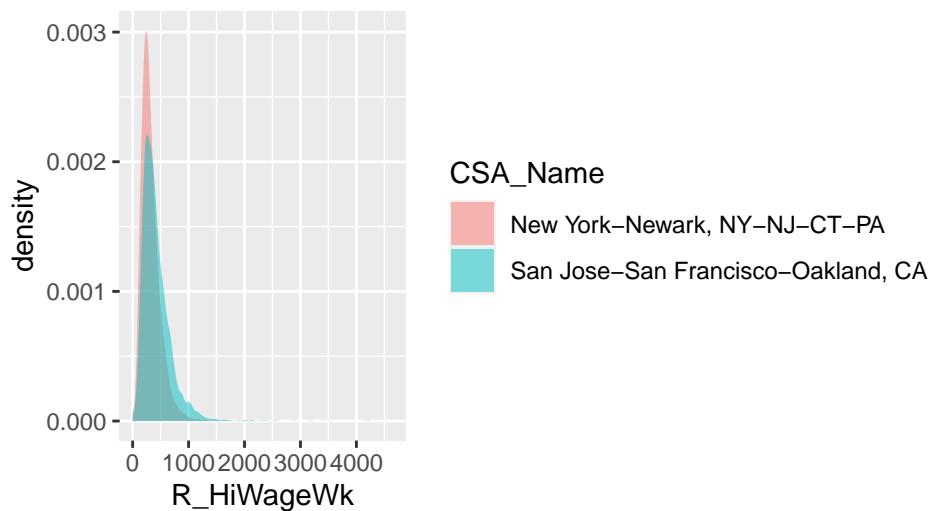
Density Plot of R_LowWageWk



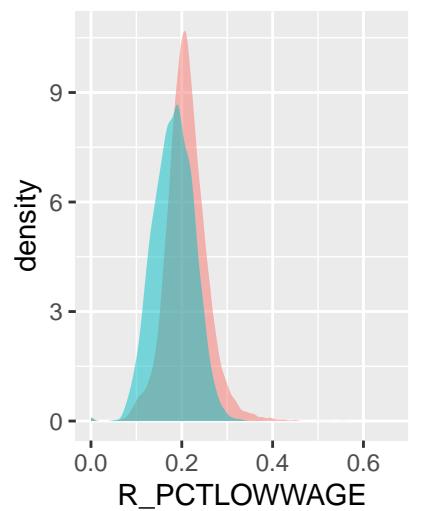
Density Plot of R_MedWageWk



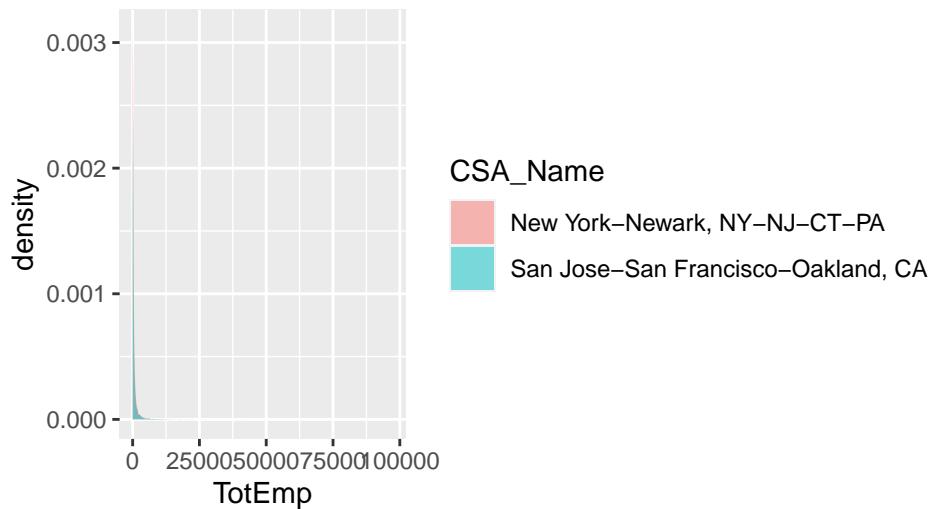
Density Plot of R_HiWageWk



Density Plot of R_PCTLLOW

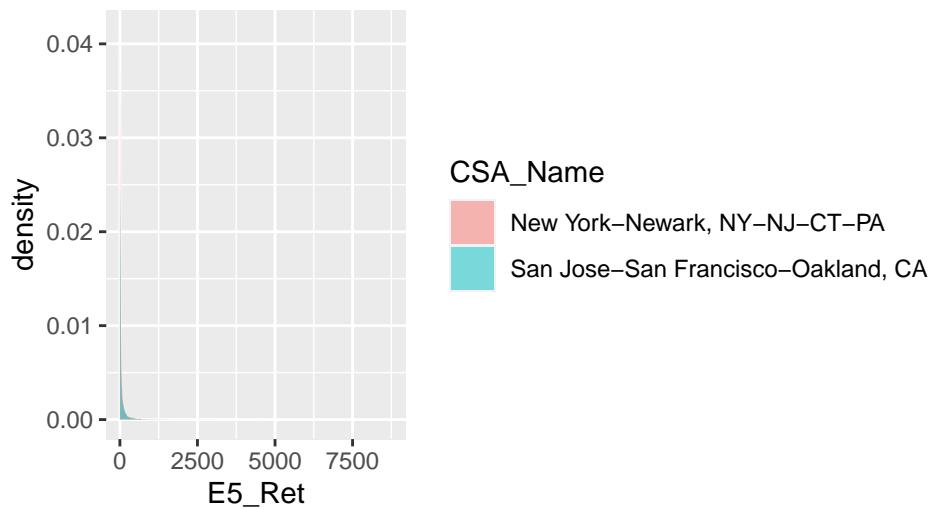


Density Plot of TotEmp

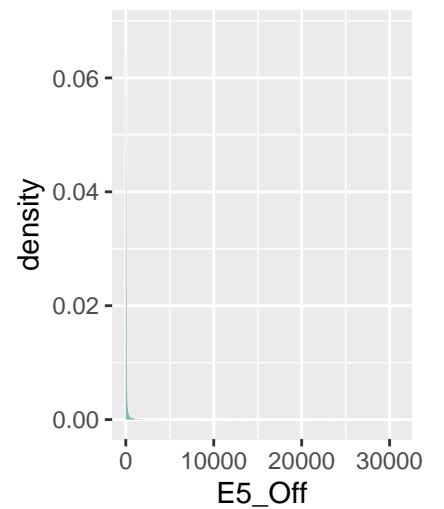


```
create_density_plots_subset(NewYork_SanFrancisco_data, 34, 54)
```

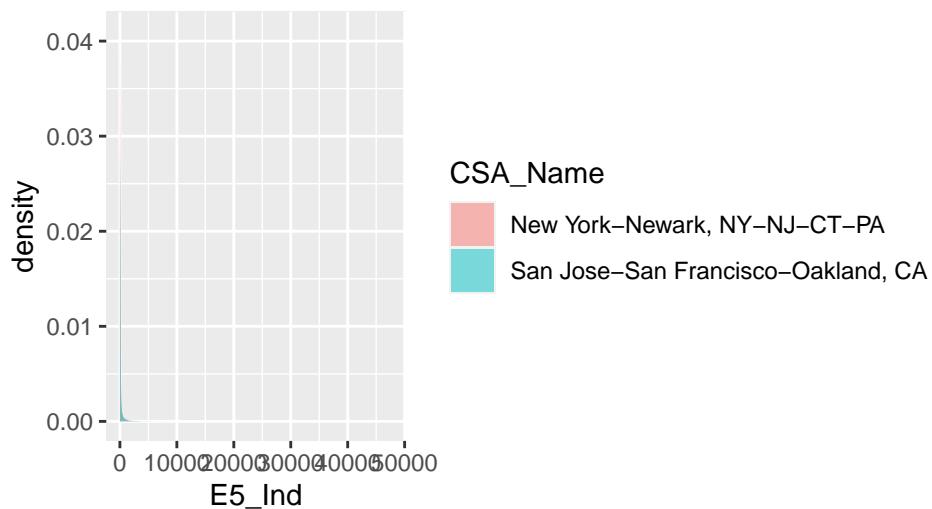
Density Plot of E5_Ret



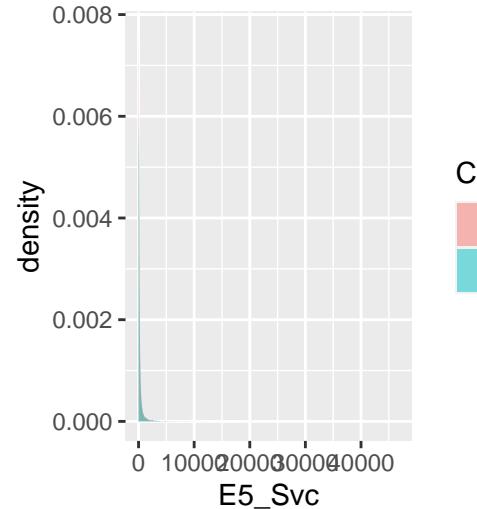
Density Plot of E5_Off



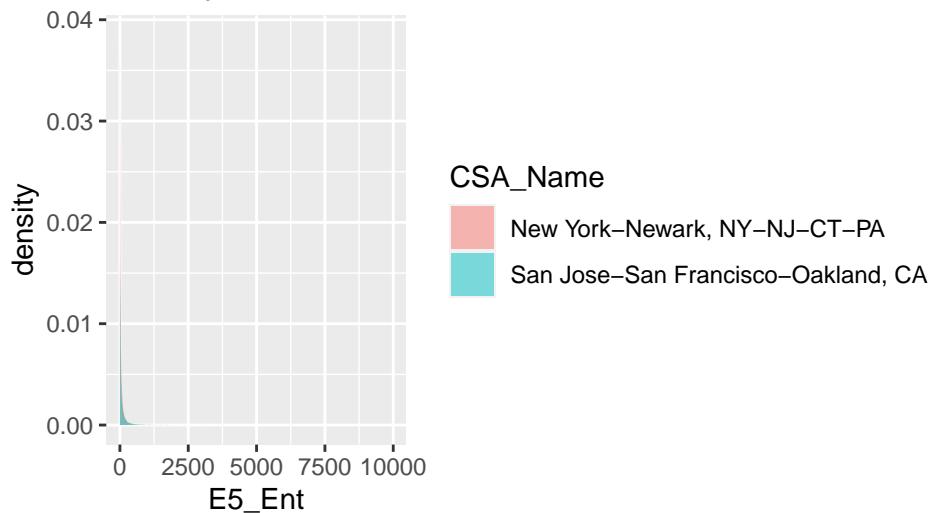
Density Plot of E5_Ind



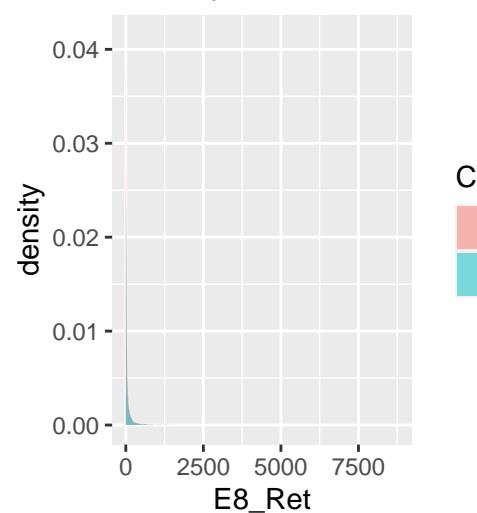
Density Plot of E5_Svc



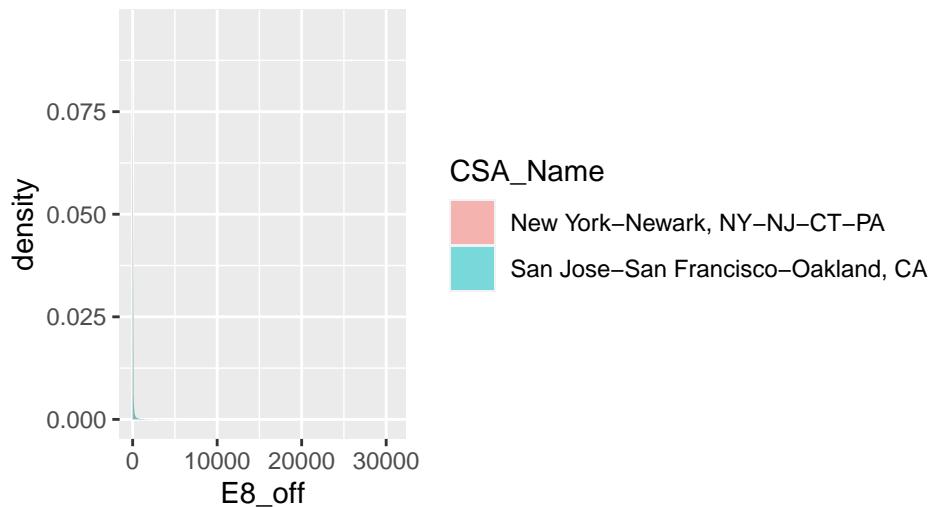
Density Plot of E5_Ent



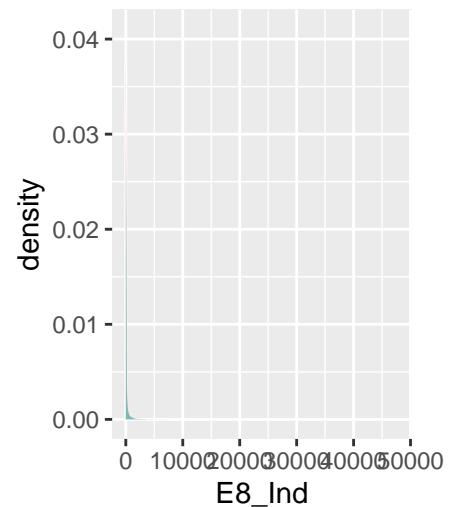
Density Plot of E8_Ret



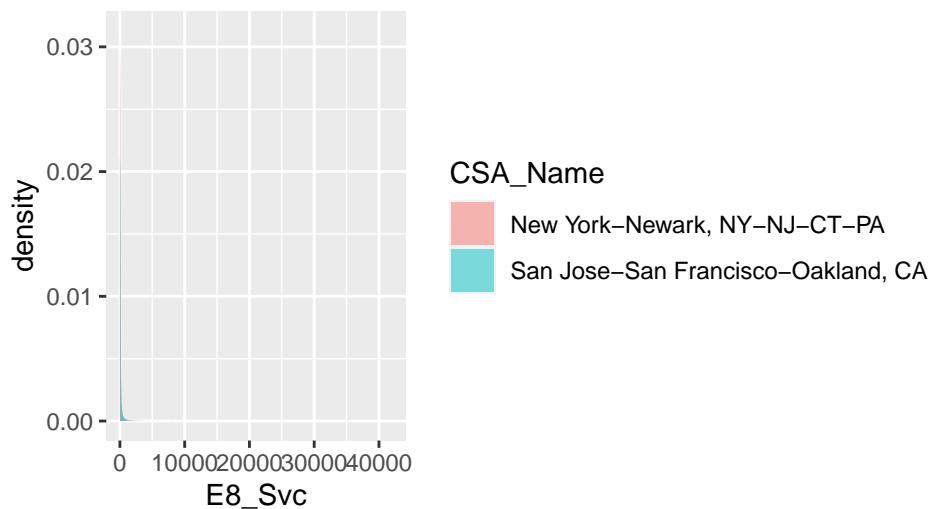
Density Plot of E8_off



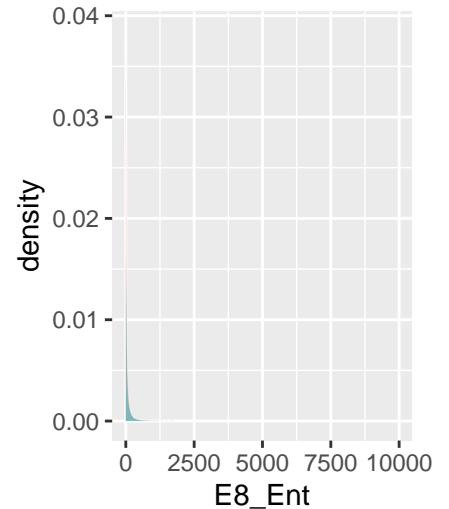
Density Plot of E8_Ind



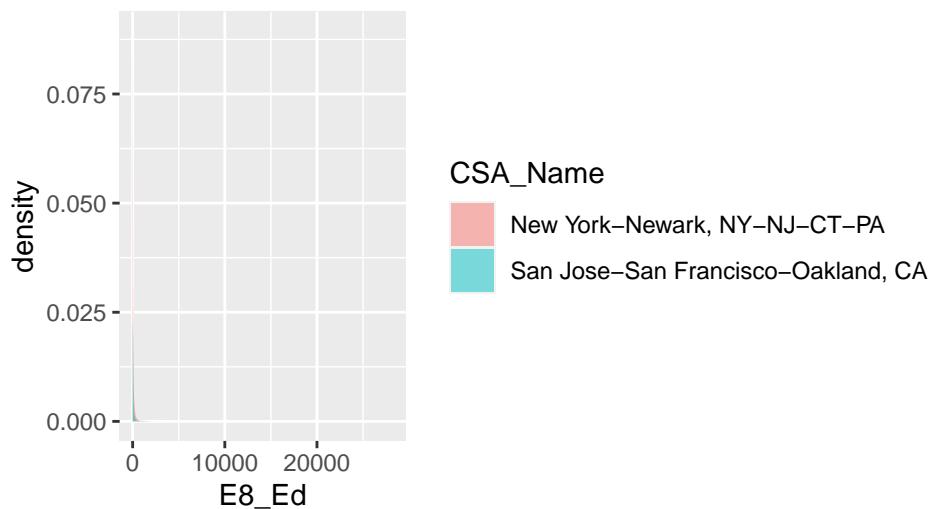
Density Plot of E8_Svc



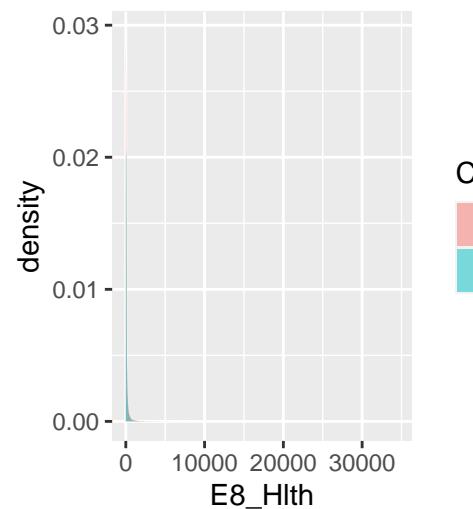
Density Plot of E8_Ent



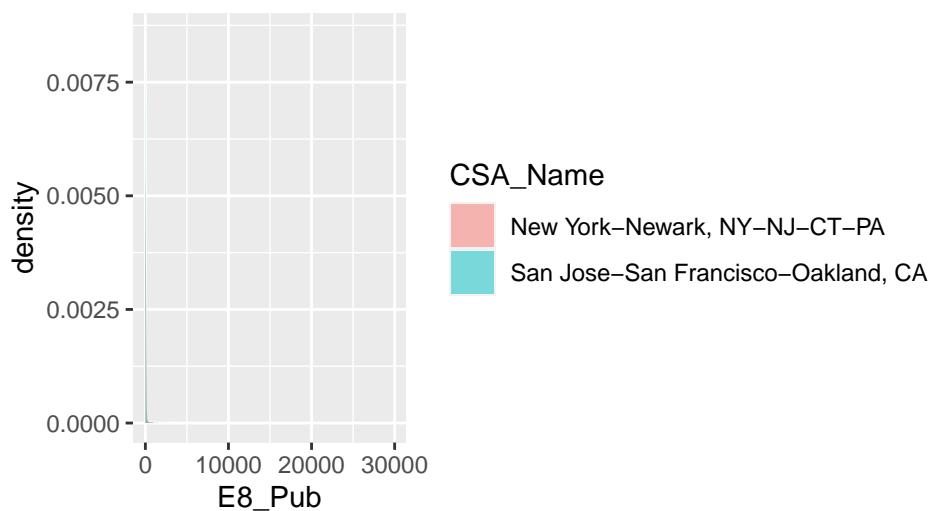
Density Plot of E8_Ed



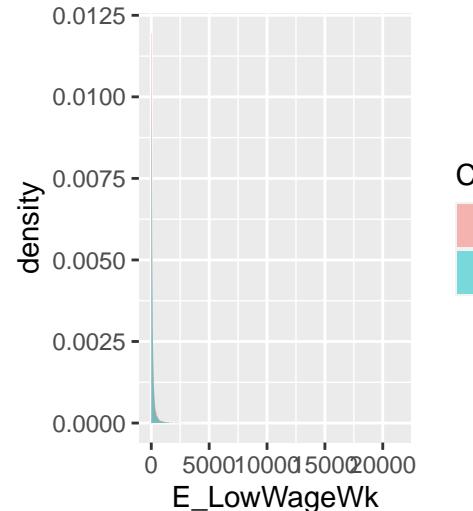
Density Plot of E8_Hlth



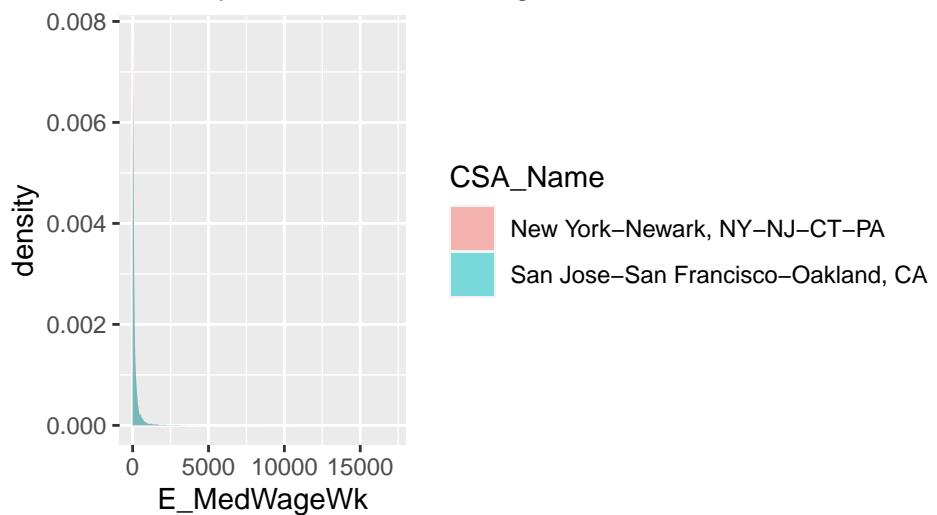
Density Plot of E8_Pub



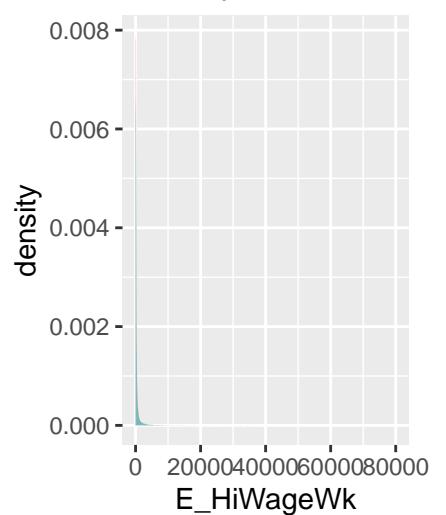
Density Plot of E_LowWageWk



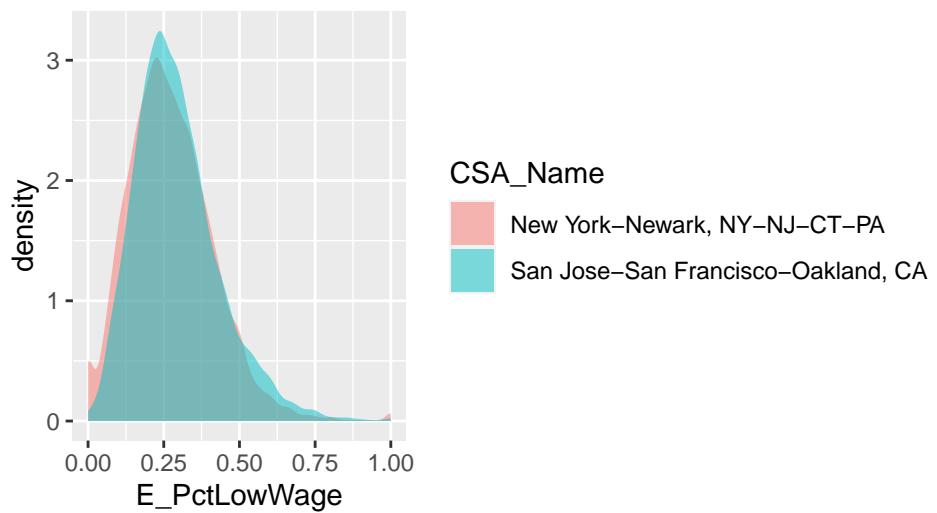
Density Plot of E_MedWageWk



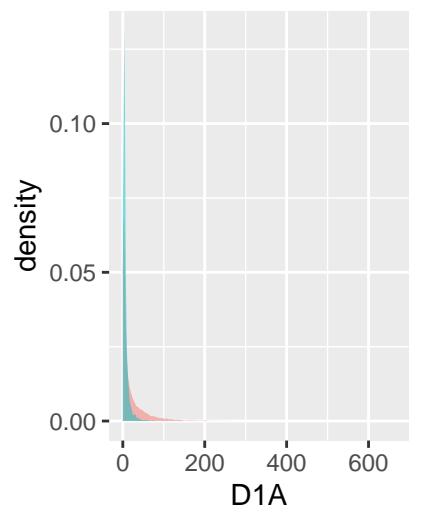
Density Plot of E_HiWageWk



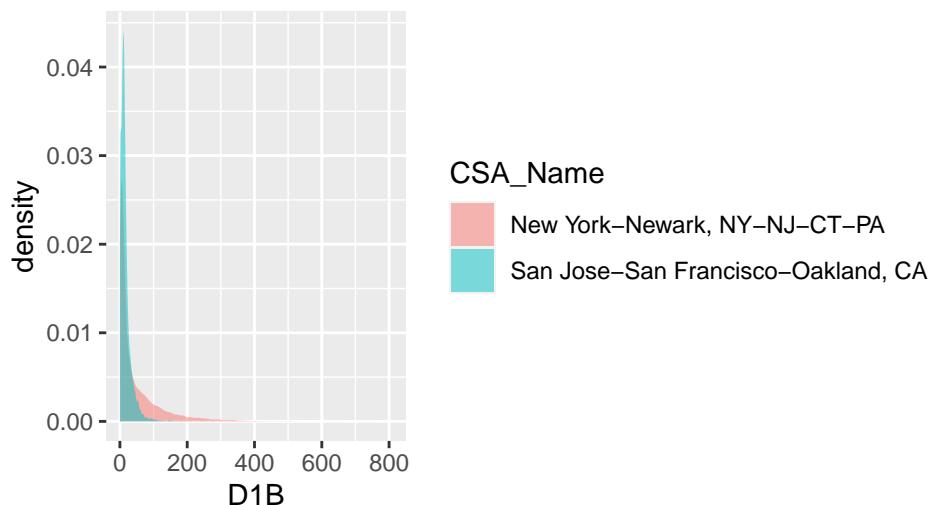
Density Plot of E_PctLowWage



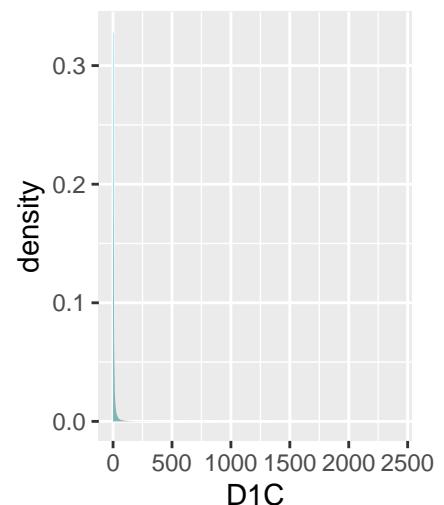
Density Plot of D1A



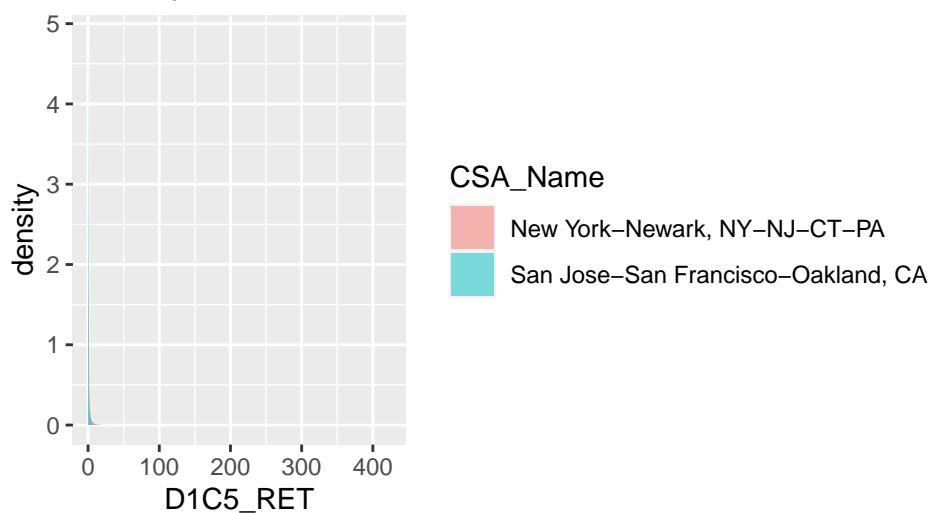
Density Plot of D1B



Density Plot of D1C

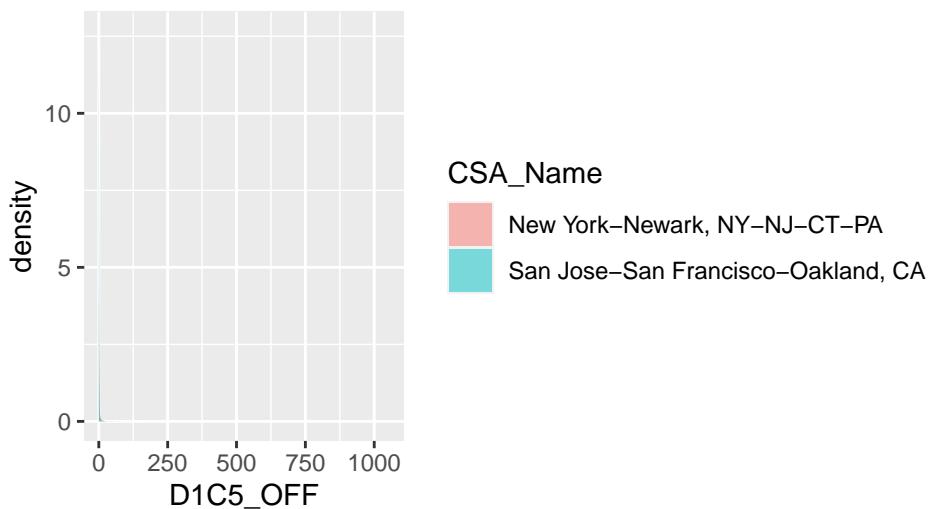


Density Plot of D1C5_RET

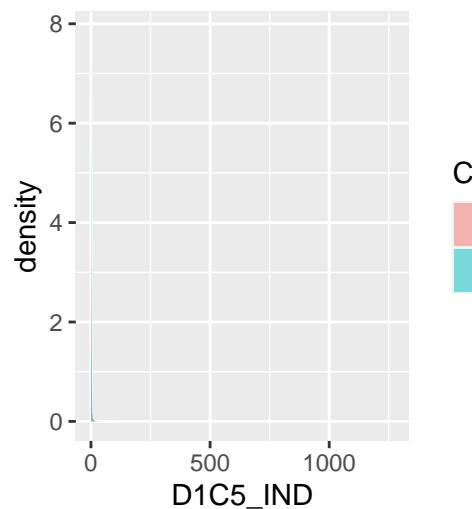


```
create_density_plots_subset(NewYork_SanFrancisco_data, 55, 75)
```

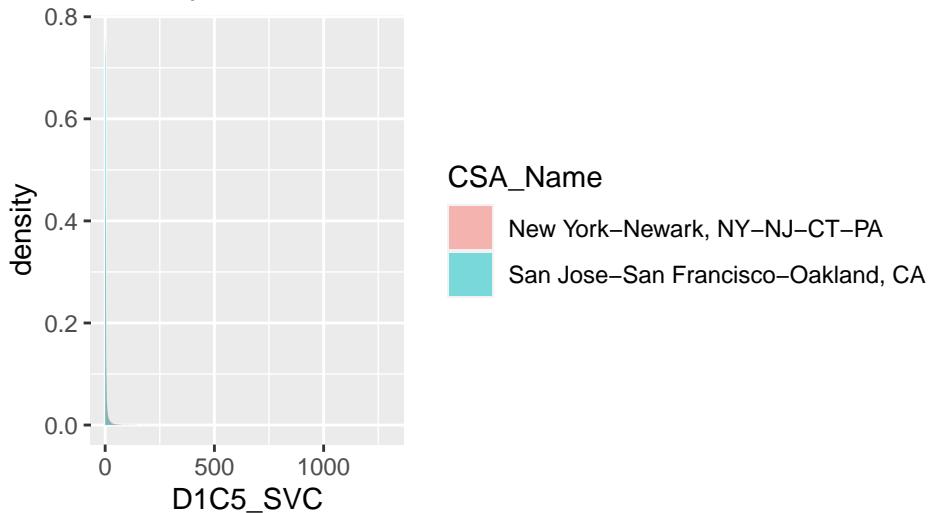
Density Plot of D1C5_OFF



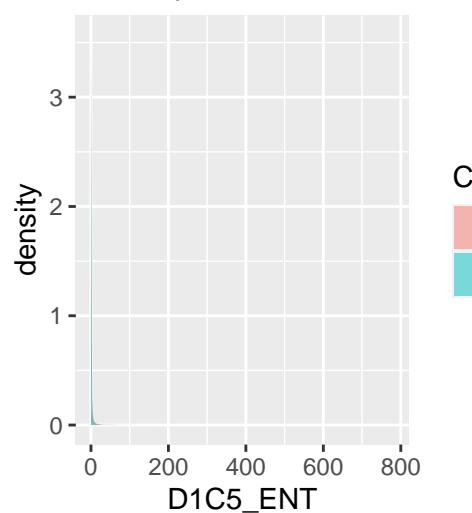
Density Plot of D1C5_IND



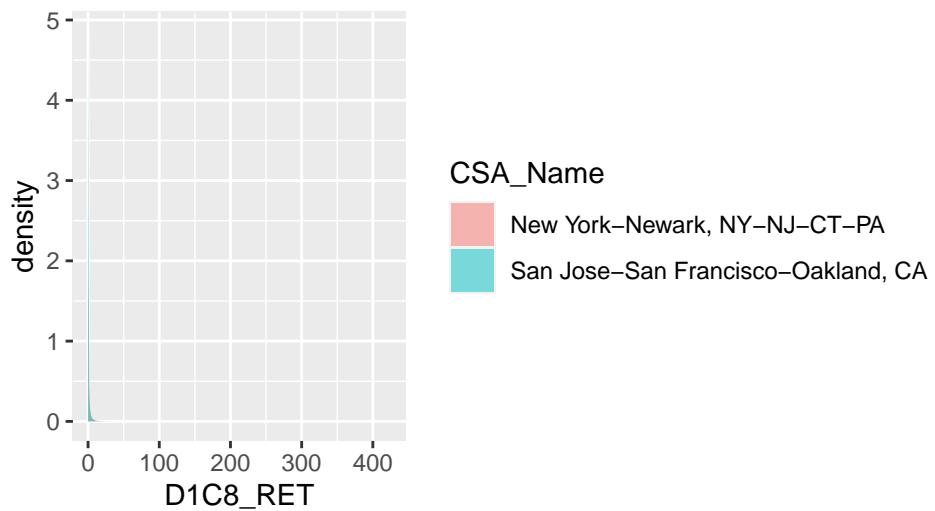
Density Plot of D1C5_SVC



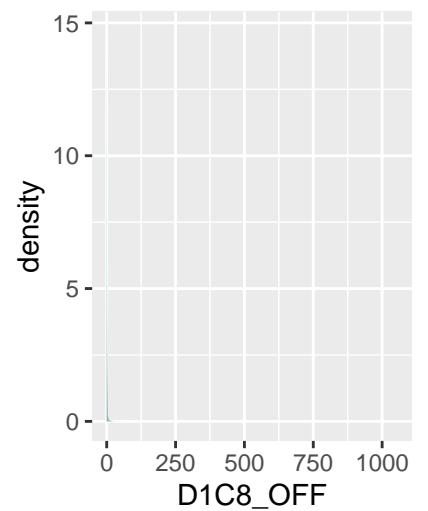
Density Plot of D1C5_ENT



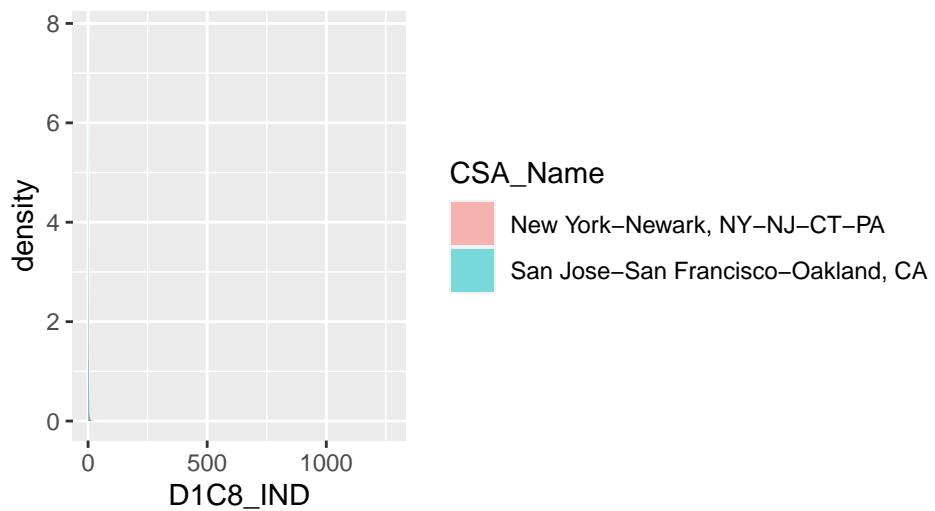
Density Plot of D1C8_RET



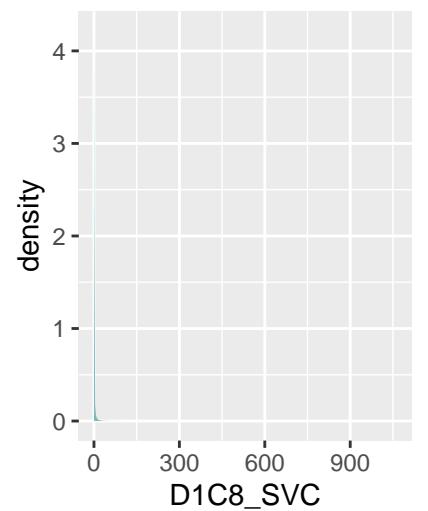
Density Plot of D1C8_OFF



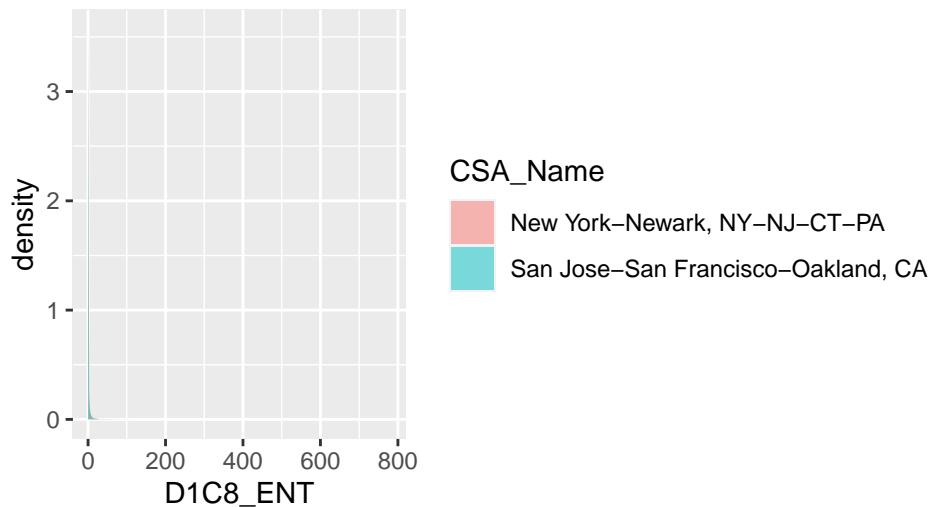
Density Plot of D1C8_IND



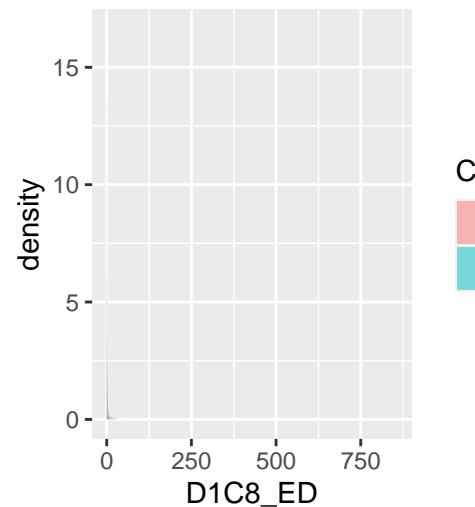
Density Plot of D1C8_SVC



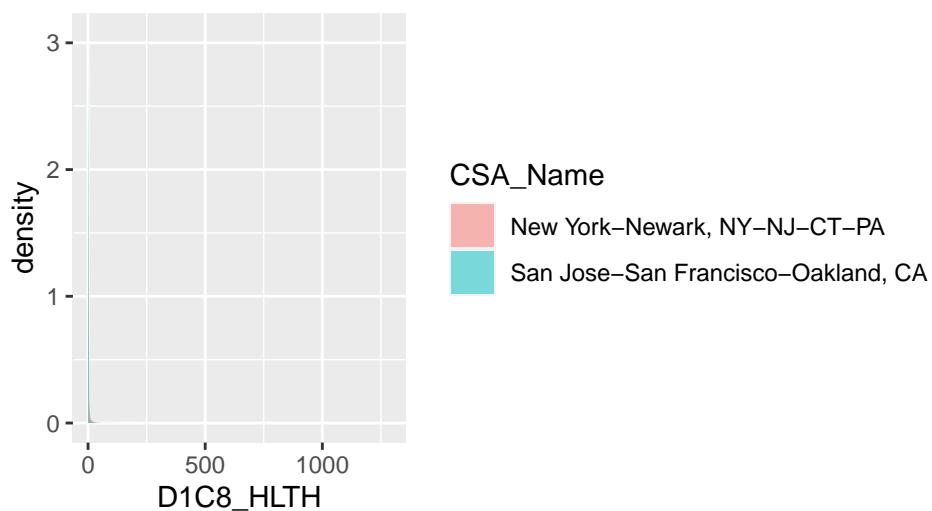
Density Plot of D1C8_ENT



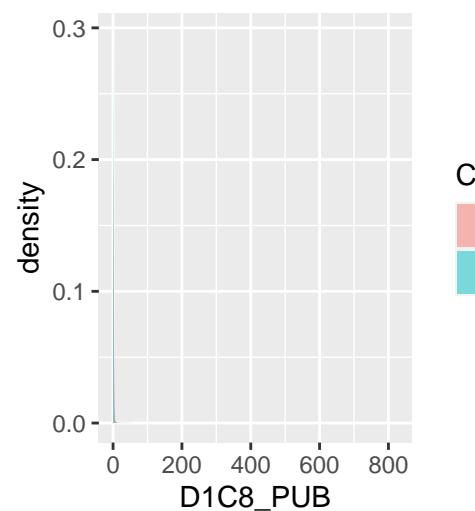
Density Plot of D1C8_ED



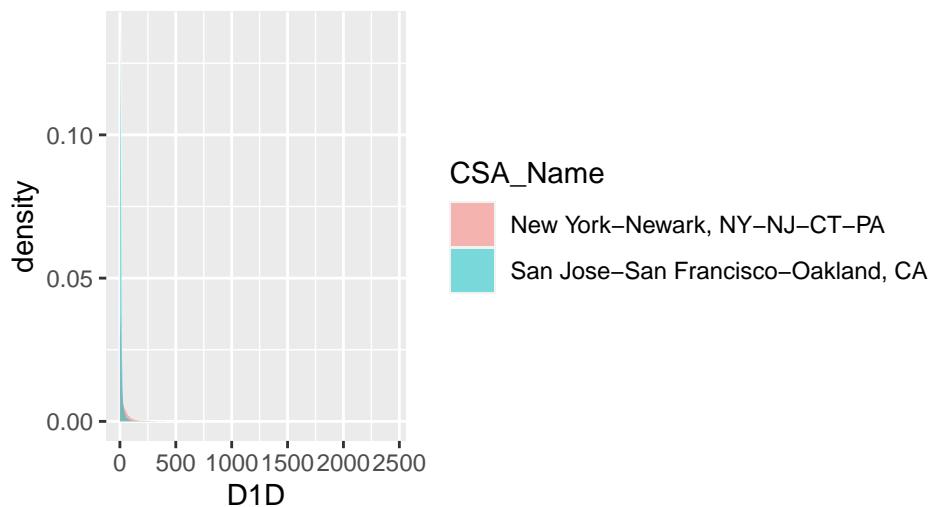
Density Plot of D1C8_HLTH



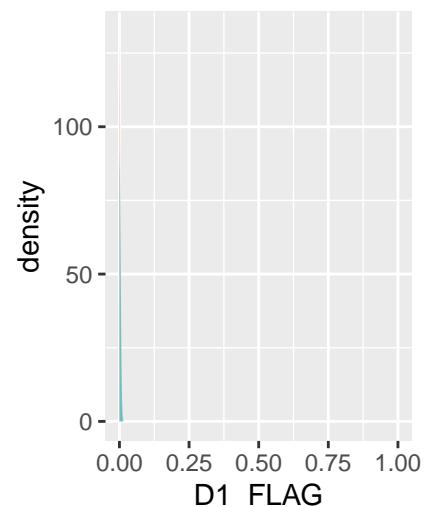
Density Plot of D1C8_PUB



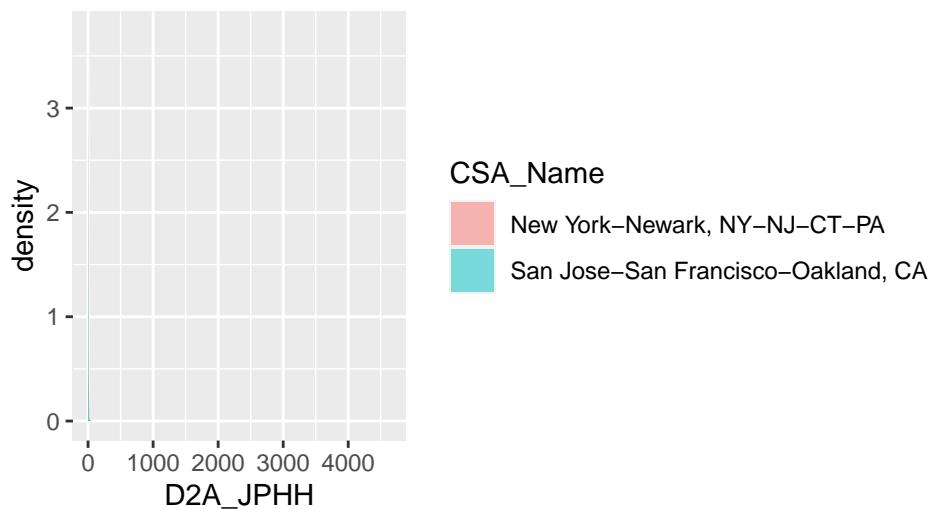
Density Plot of D1D



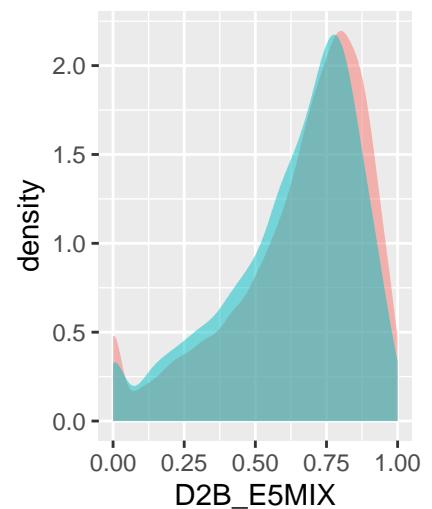
Density Plot of D1_FLAG



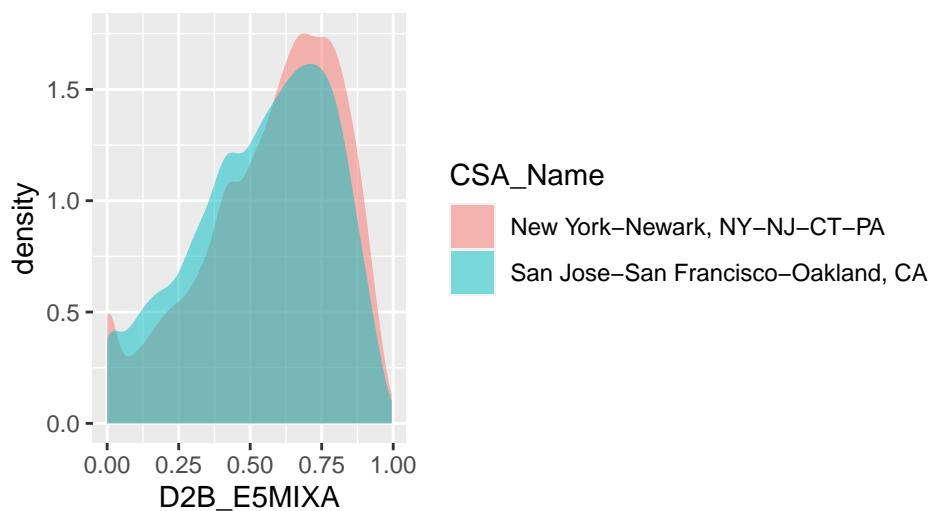
Density Plot of D2A_JPHH



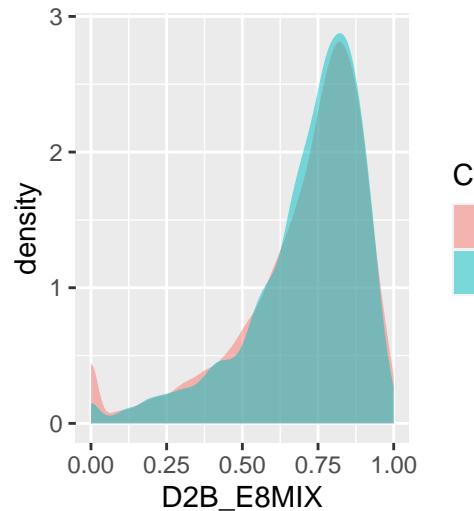
Density Plot of D2B_E5MIX



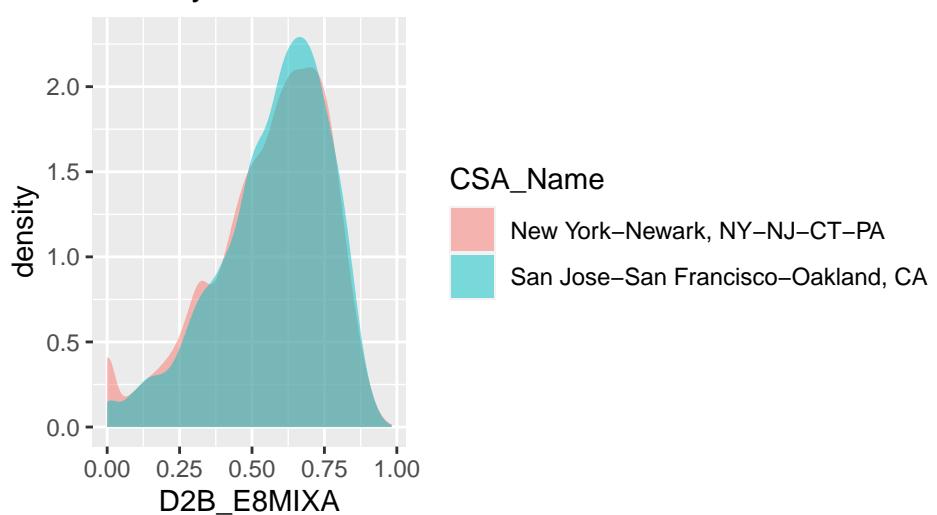
Density Plot of D2B_E5MIXA



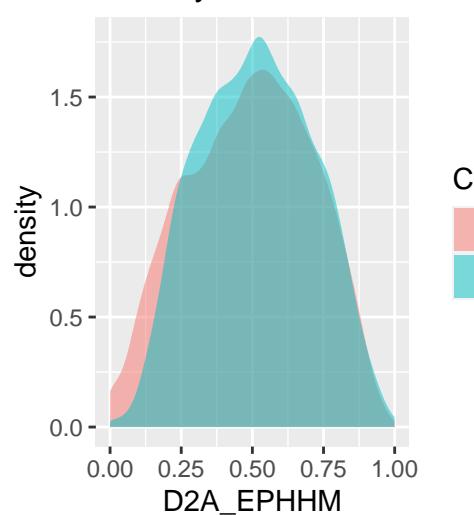
Density Plot of D2B_E8MIXA



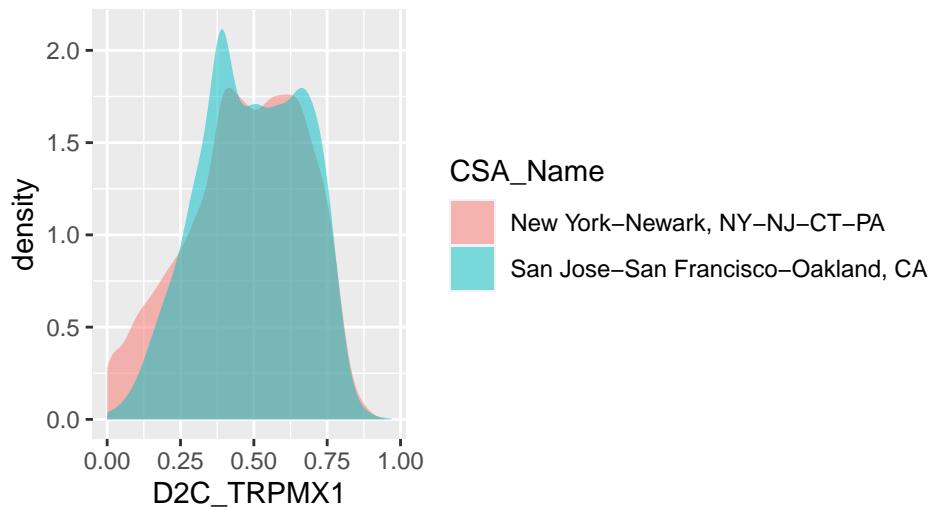
Density Plot of D2B_E8MIXA



Density Plot of D2A_EPHHM

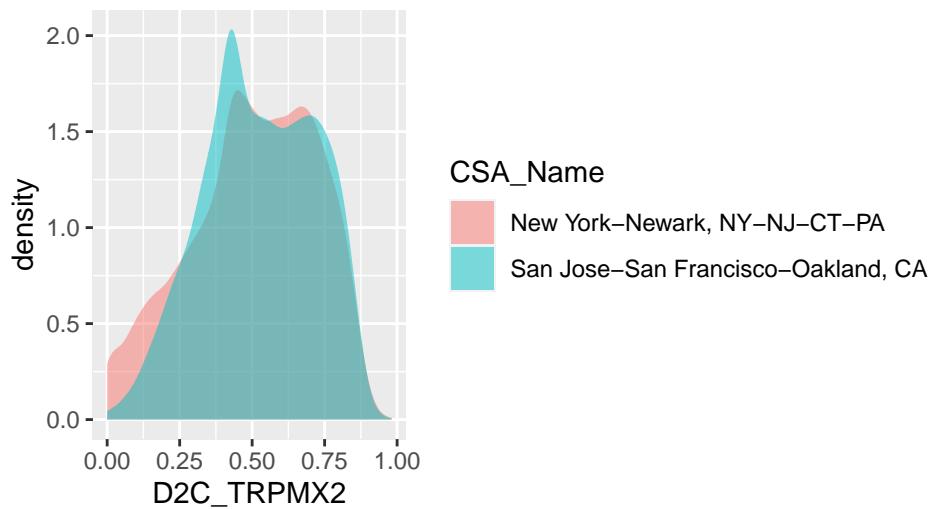


Density Plot of D2C_TRPMX1

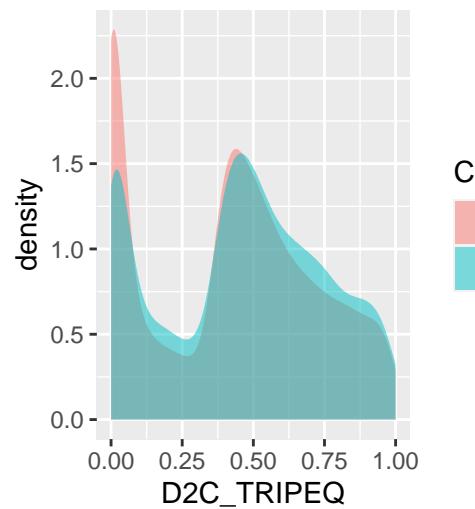


```
create_density_plots_subset(NewYork_SanFrancisco_data, 76, 96)
```

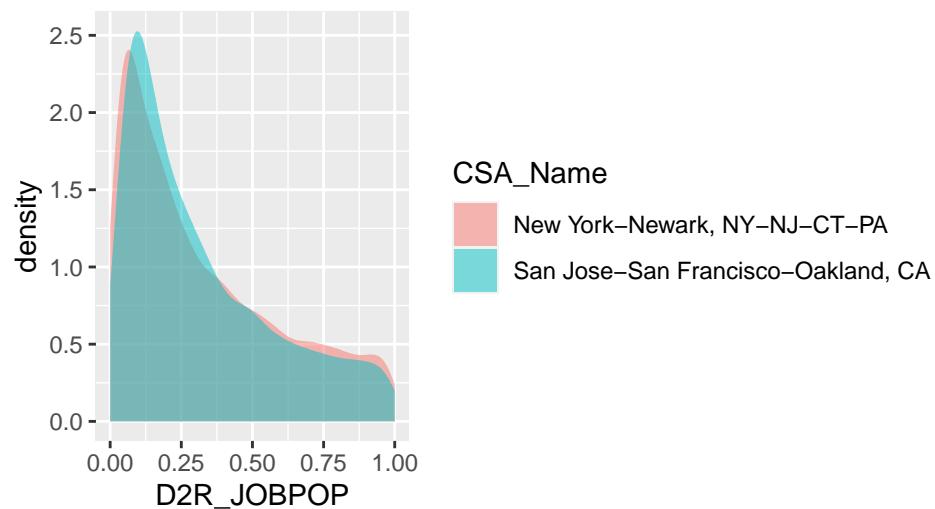
Density Plot of D2C_TRPMX2



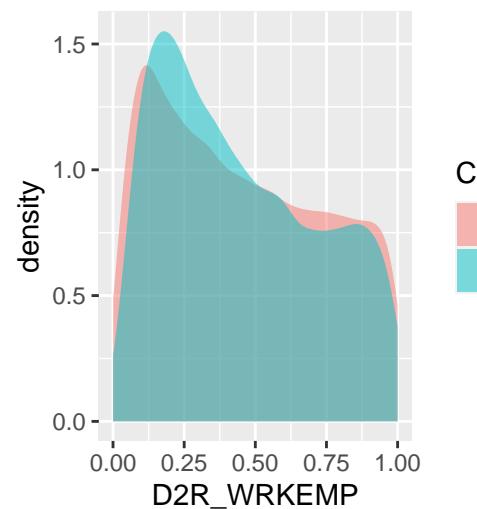
Density Plot of D2C_TRIPEQ



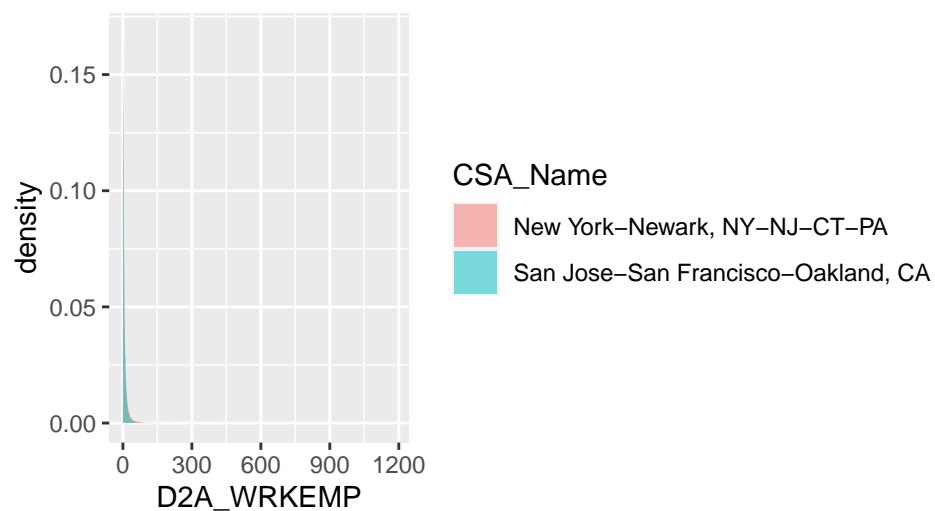
Density Plot of D2R_JOBPOP



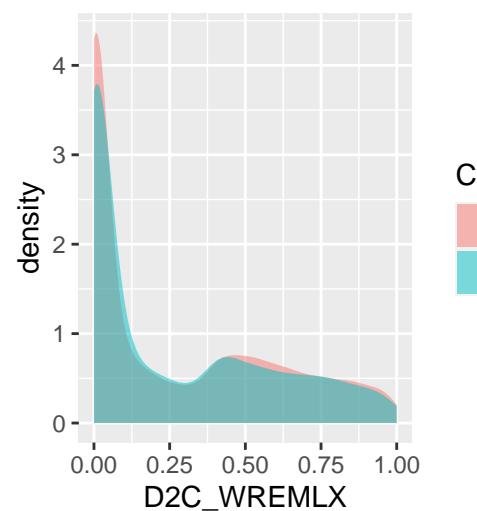
Density Plot of D2R_WRKEMP



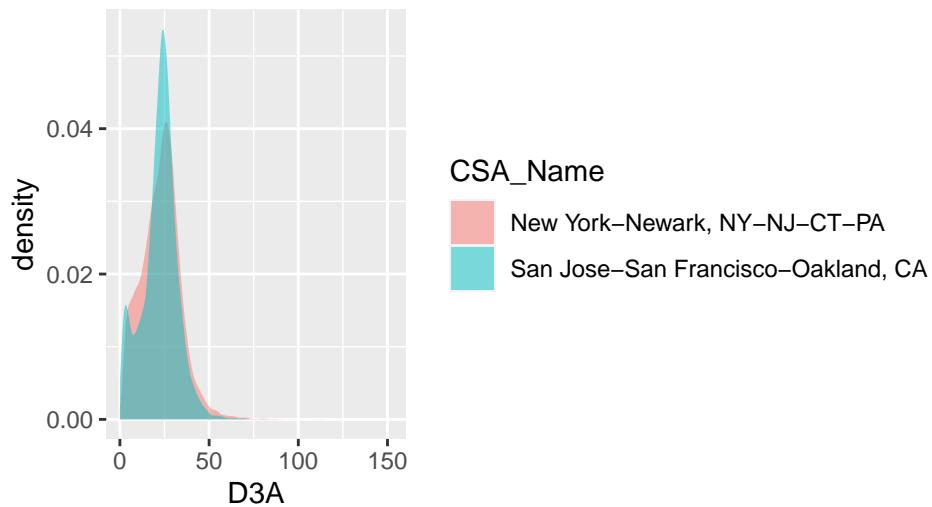
Density Plot of D2A_WRKEMP



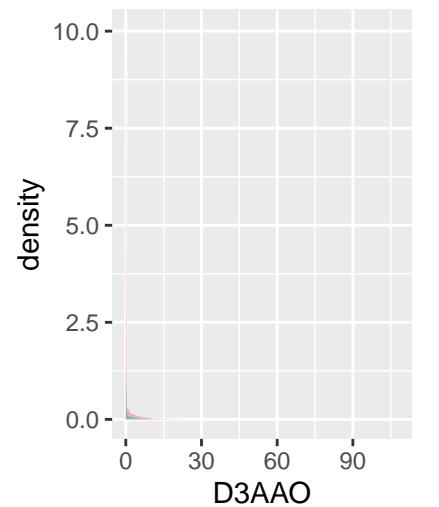
Density Plot of D2C_WREMLX



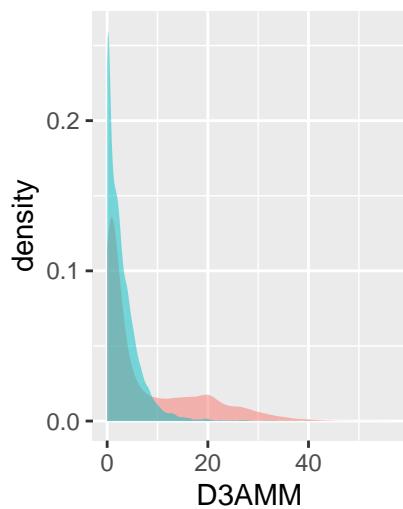
Density Plot of D3A



Density Plot of D3AAO



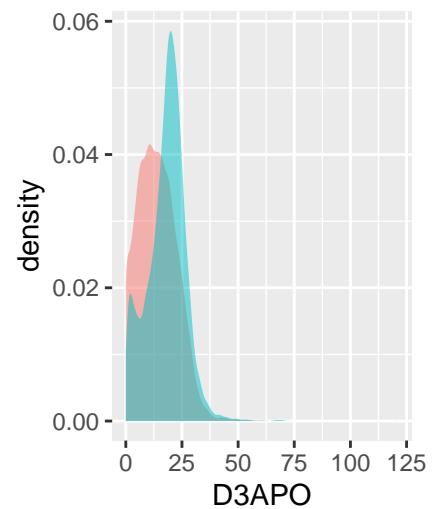
Density Plot of D3AMM



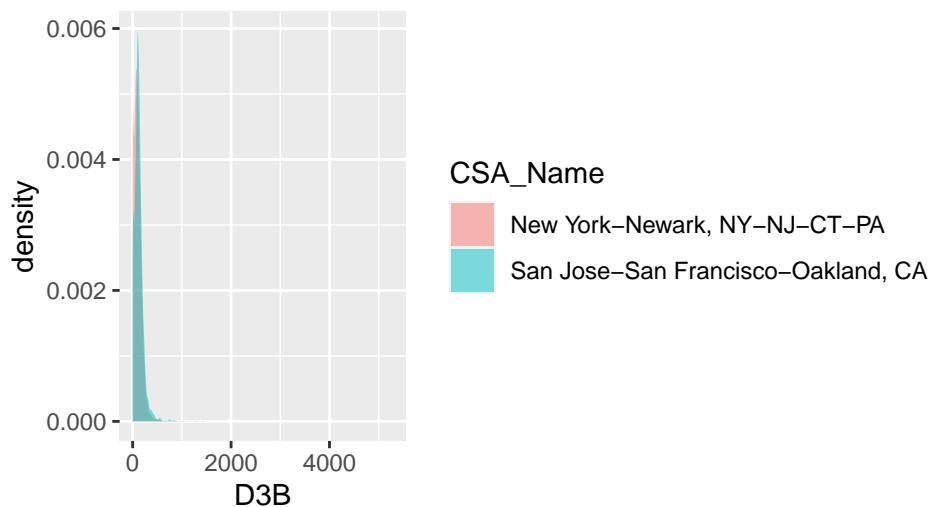
CSA_Name

- New York-Newark, NY-NJ-CT-PA
- San Jose-San Francisco-Oakland, CA

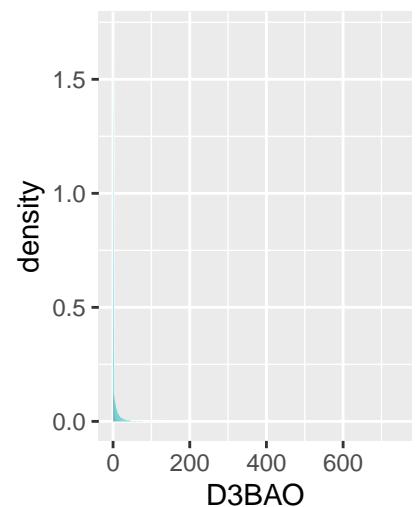
Density Plot of D3APO



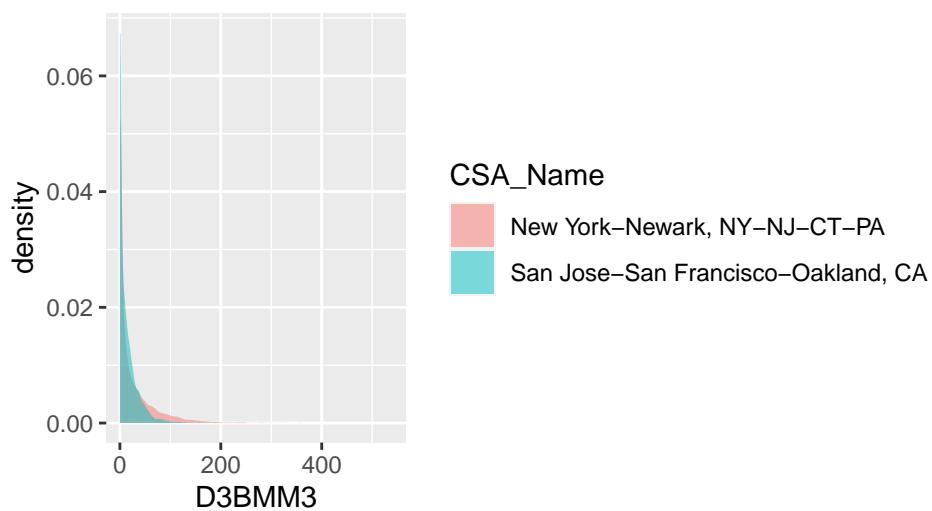
Density Plot of D3B



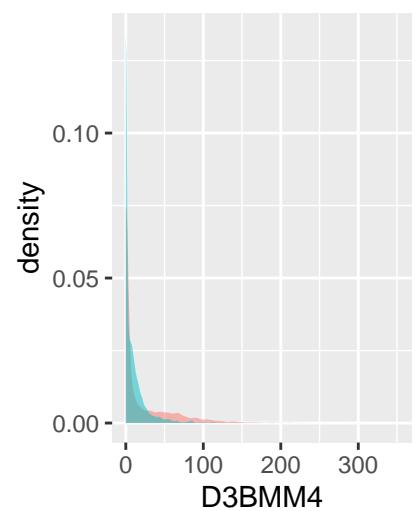
Density Plot of D3BAO



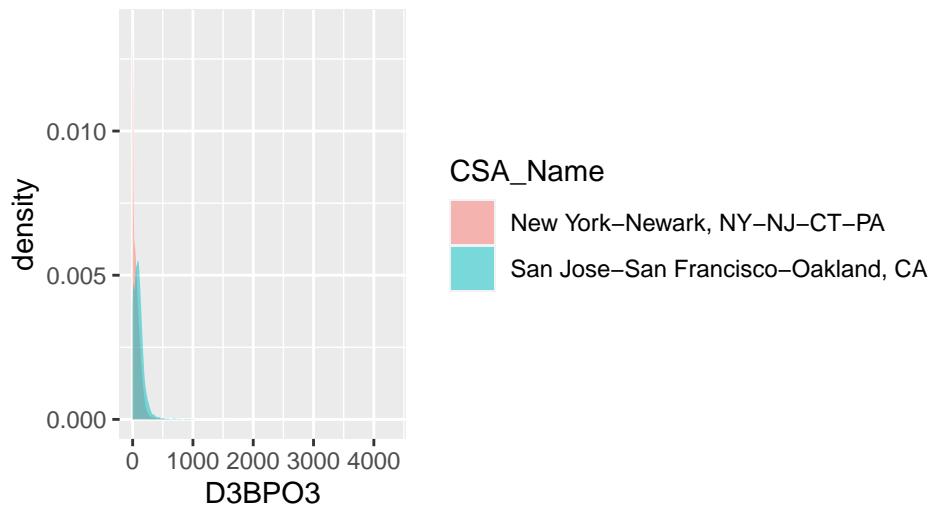
Density Plot of D3BMM3



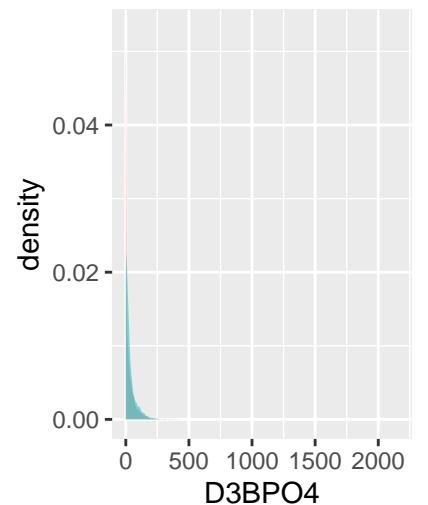
Density Plot of D3BMM4



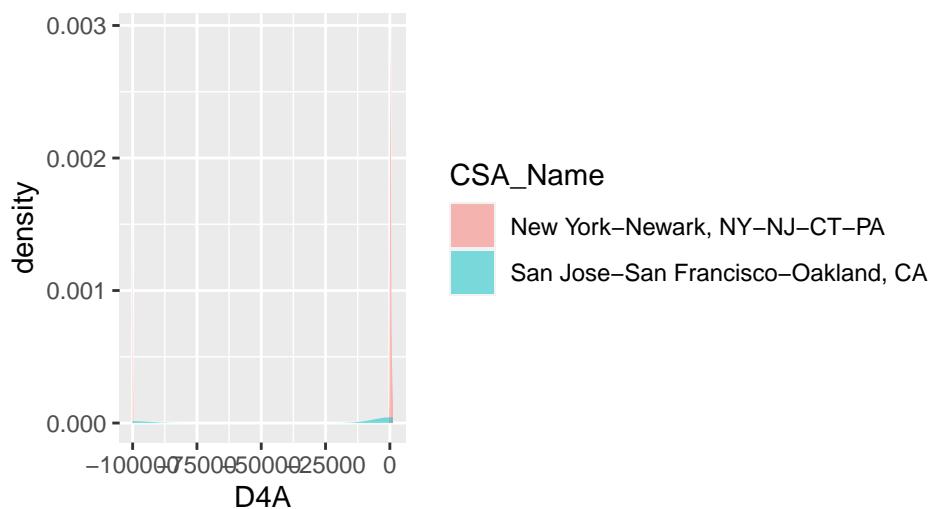
Density Plot of D3BPO3



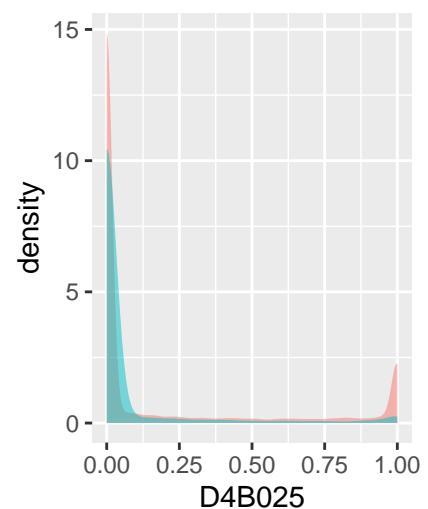
Density Plot of D3BPO4



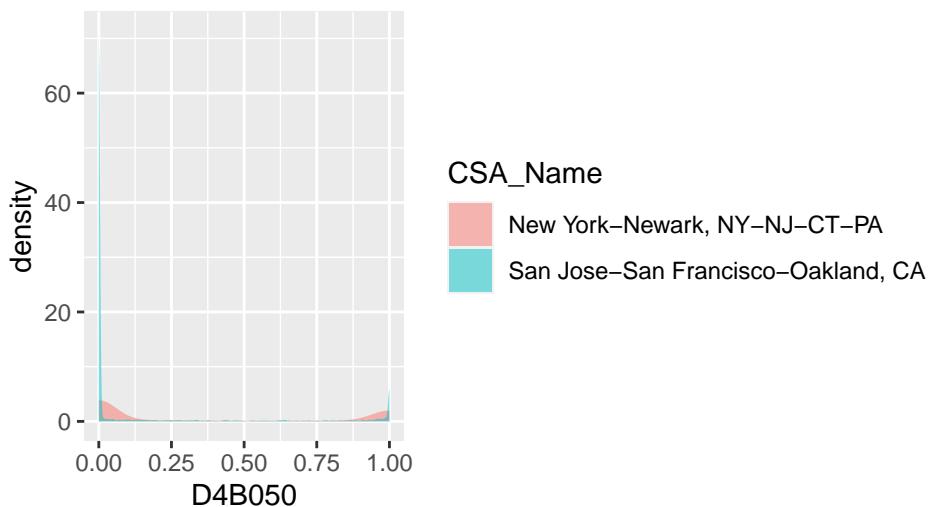
Density Plot of D4A



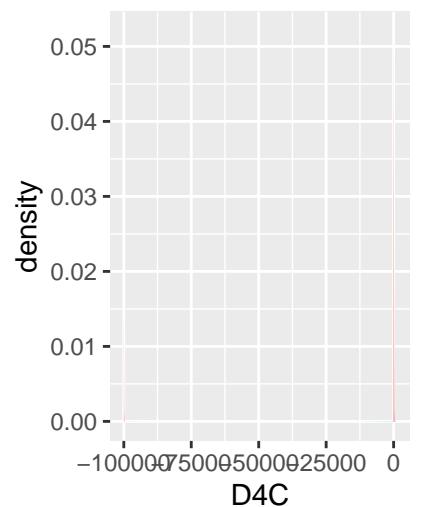
Density Plot of D4B025



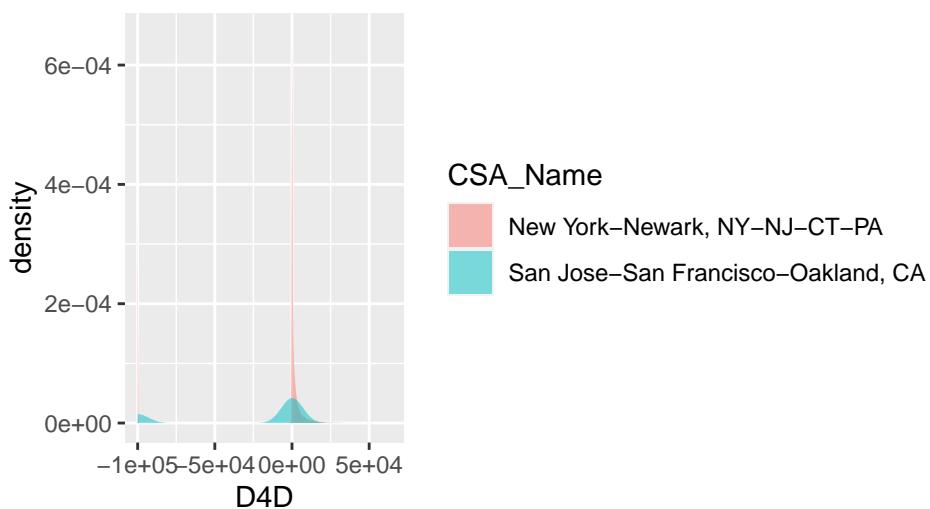
Density Plot of D4B050



Density Plot of D4C

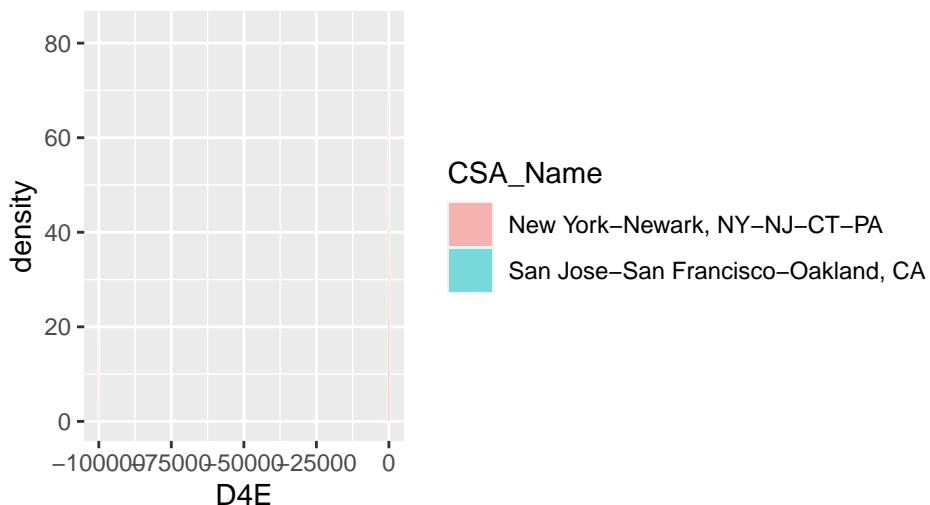


Density Plot of D4D

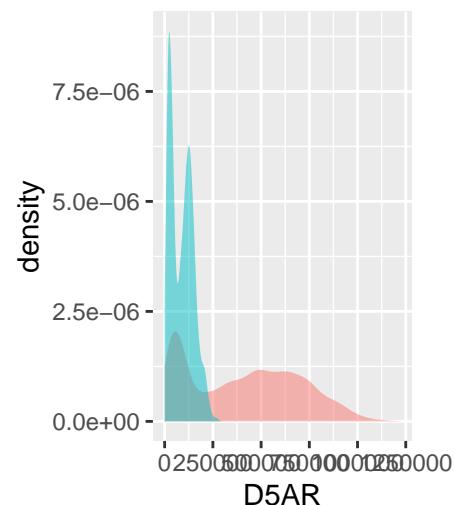


```
create_density_plots_subset(NewYork_SanFrancisco_data, 97, 118)
```

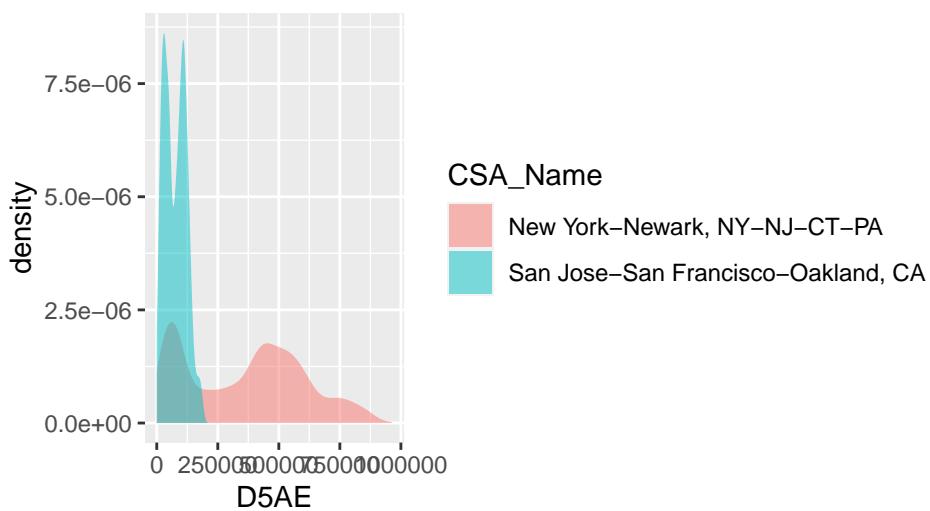
Density Plot of D4E



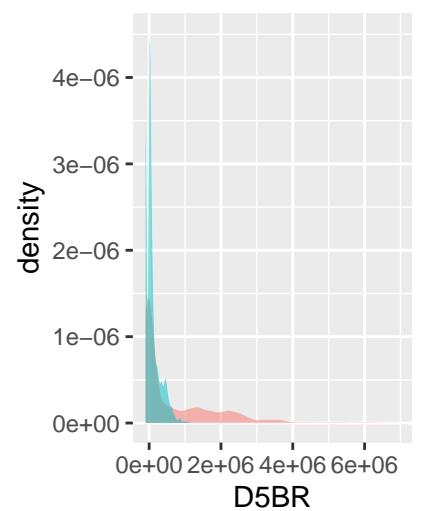
Density Plot of D5AR

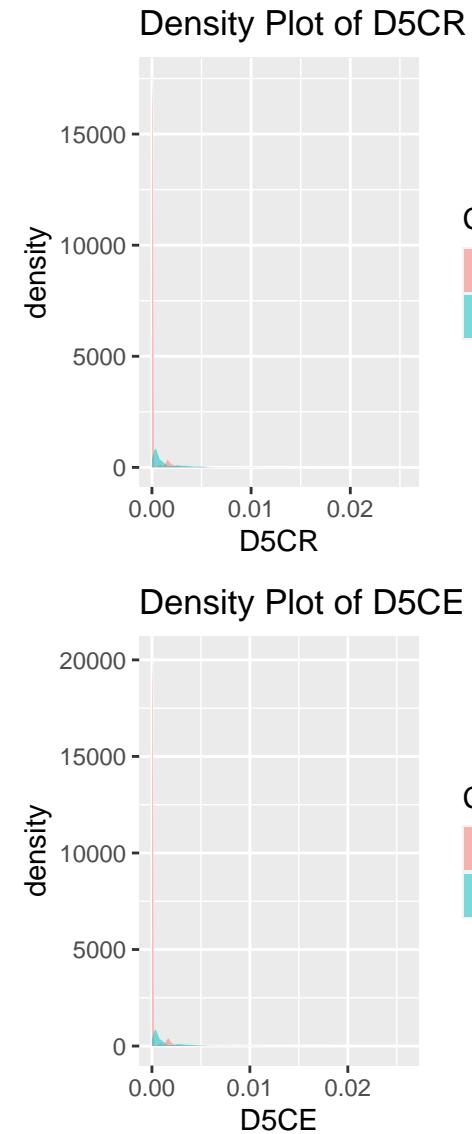
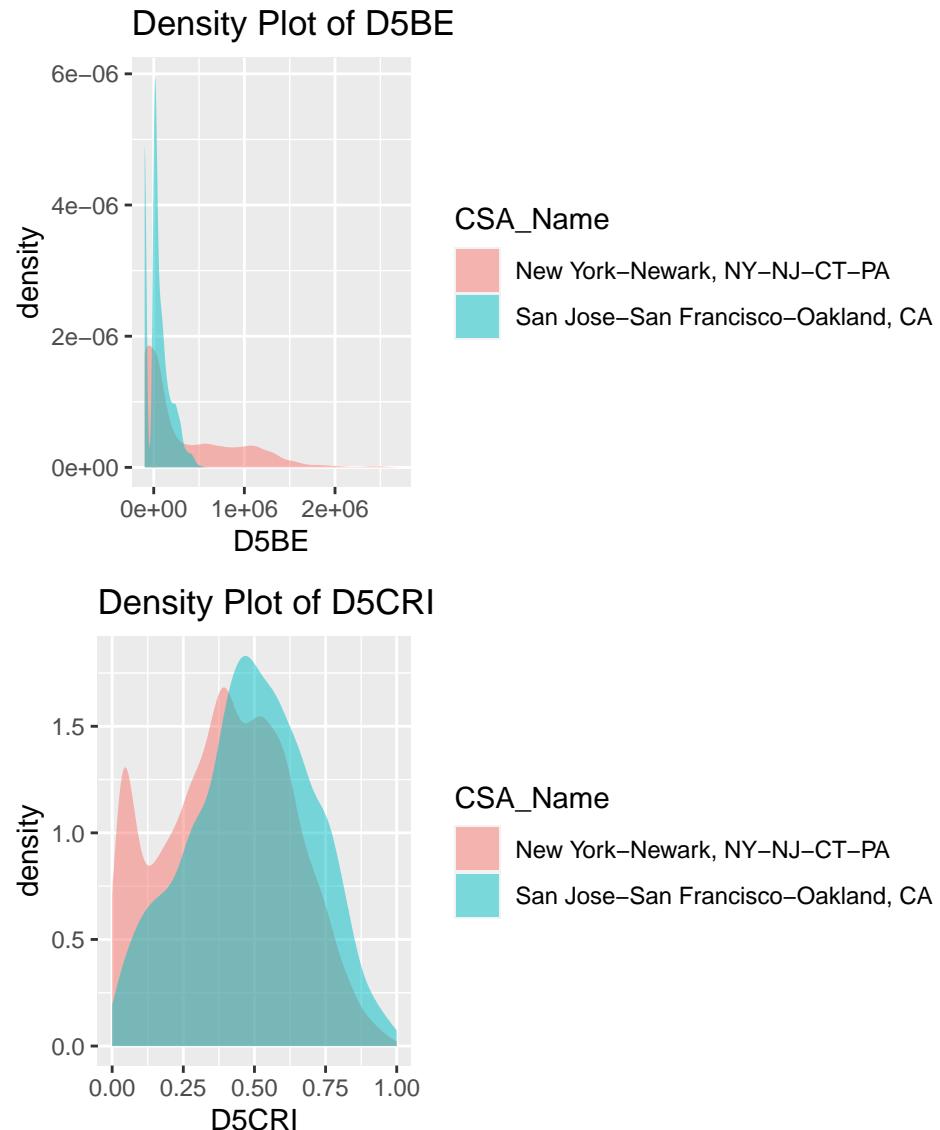


Density Plot of D5AE

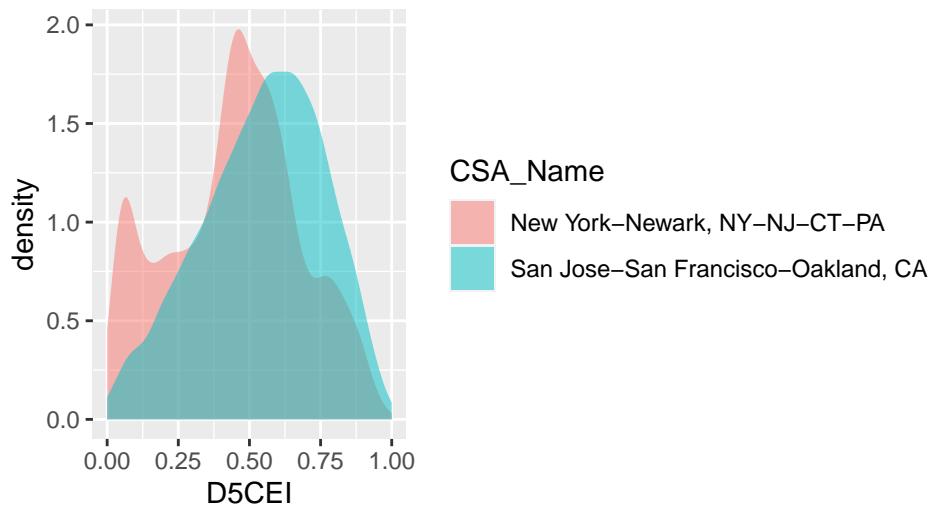


Density Plot of D5BR

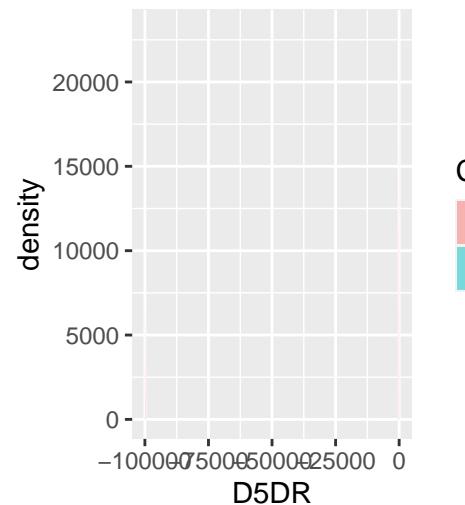




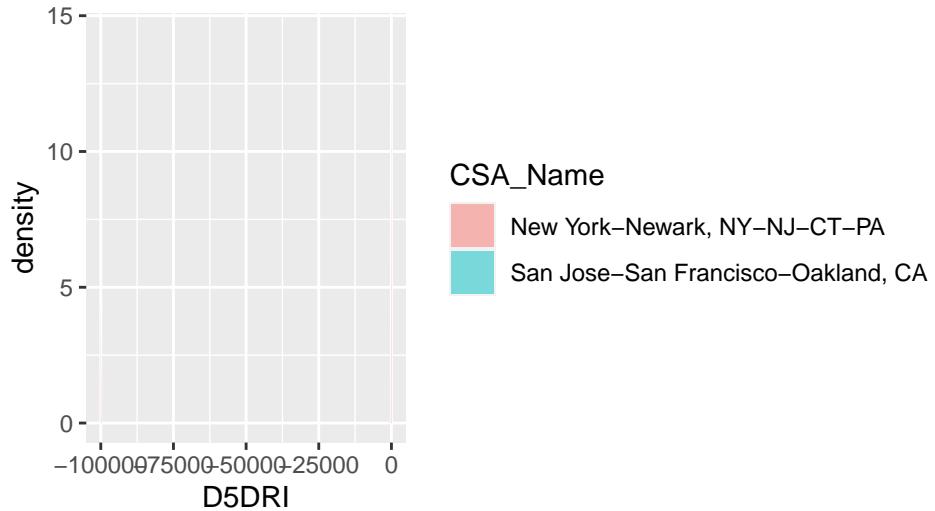
Density Plot of D5CEI



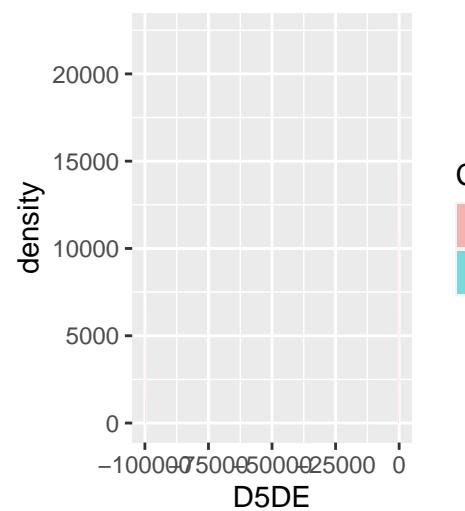
Density Plot of D5DR



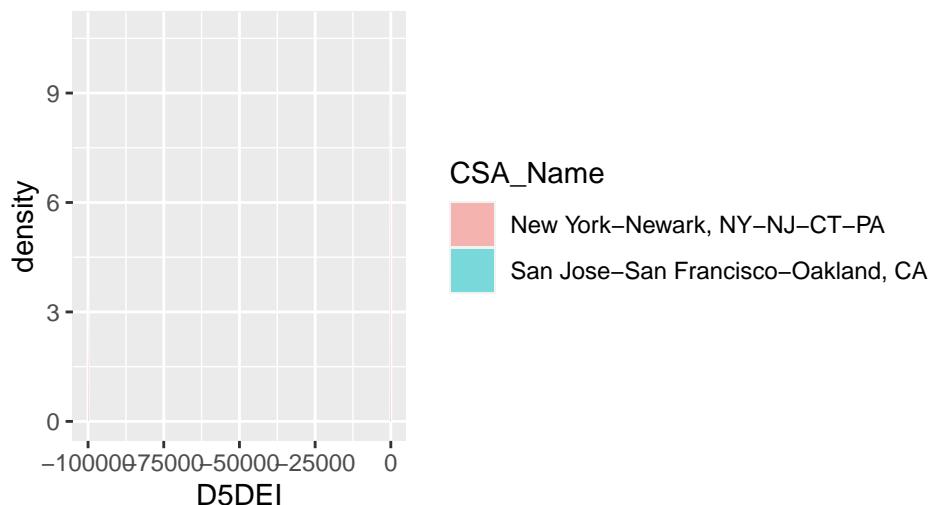
Density Plot of D5DRI



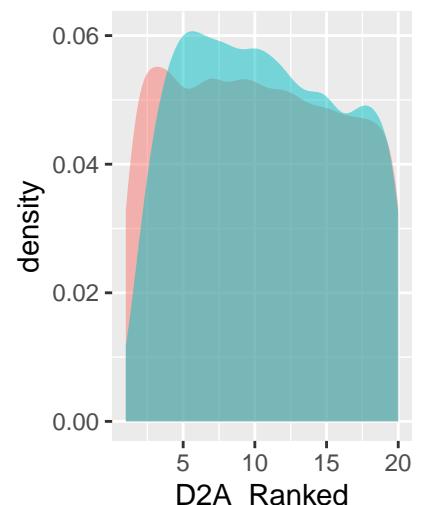
Density Plot of D5DE



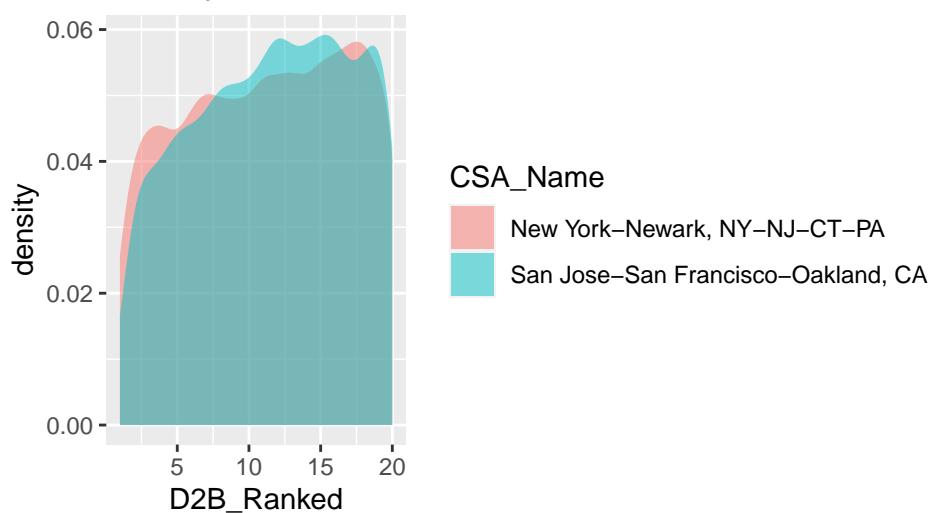
Density Plot of D5DEI



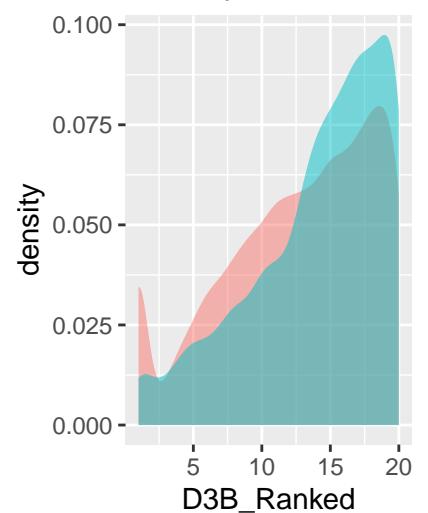
Density Plot of D2A_Ranked



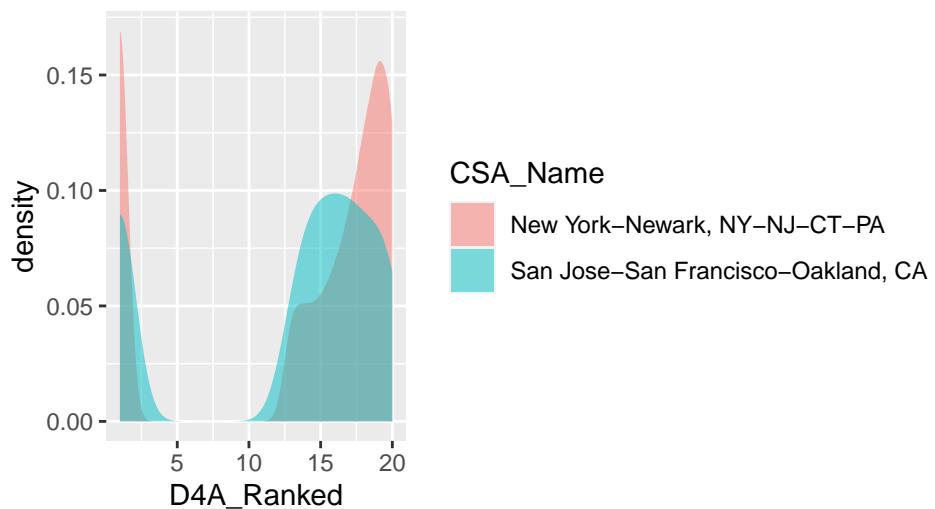
Density Plot of D2B_Ranked



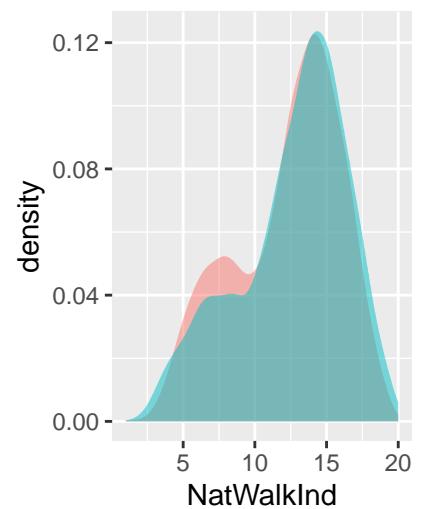
Density Plot of D3B_Ranked



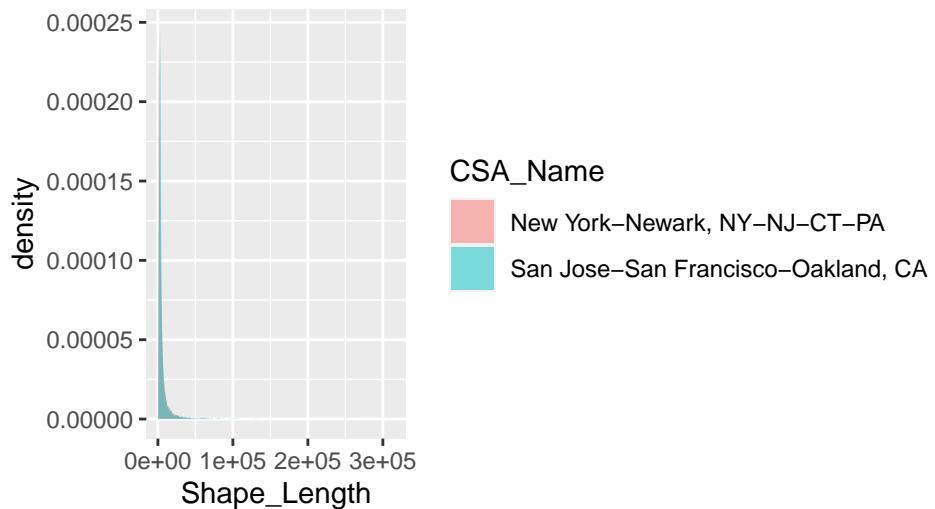
Density Plot of D4A_Ranked



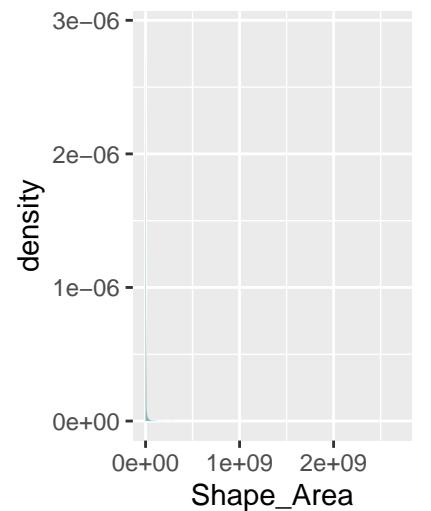
Density Plot of NatWalkInd

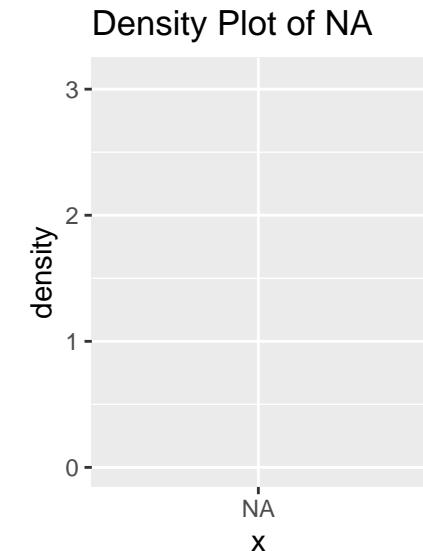
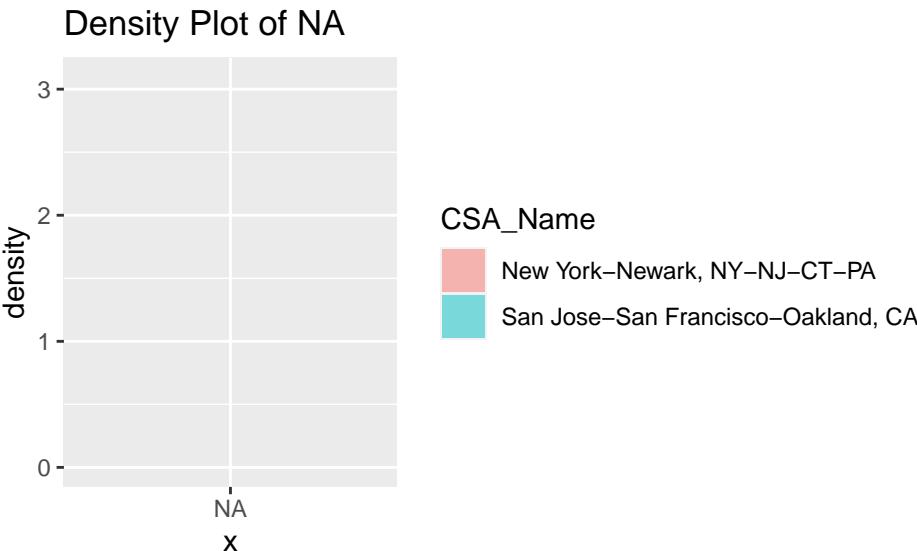


Density Plot of Shape_Length



Density Plot of Shape_Area





For columns 11 to 20:

```
#Selecting Variables (there is way too many of them to run ggpairs on all of them)
```

```
NewYork_SanFrancisco_data <- NewYork_SanFrancisco_data %>%
  dplyr::mutate(CSA_Name_ABBREV = gsub("New York-Newark, NY-NJ-CT-PA", "NY", CSA_Name)) %>%
  mutate(CSA_Name_ABBREV = gsub("San Jose-San Francisco-Oakland, CA", "SF", CSA_Name))

# Remove leading and trailing whitespace
NewYork_SanFrancisco_data$CSA_Name_ABBREV <- trimws(NewYork_SanFrancisco_data$CSA_Name_ABBREV)

# Replace "New York-Newark, NY-NJ-CT-PA" with "NY"
NewYork_SanFrancisco_data$CSA_Name_ABBREV <- gsub("New York-Newark, NY-NJ-CT-PA", "NY", NewYork_SanFrancisco_data$CSA_Name_ABBREV)

# Replace "San Jose-San Francisco-Oakland, CA" with "SF"
NewYork_SanFrancisco_data$CSA_Name_ABBREV <- gsub("San Jose-San Francisco-Oakland, CA", "SF", NewYork_SanFrancisco_data$CSA_Name_ABBREV)

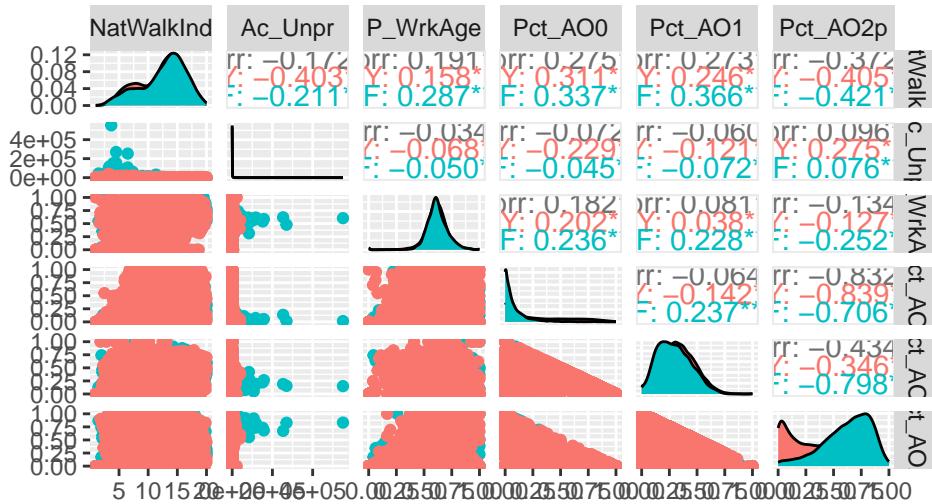
NY_SF_focus <- NewYork_SanFrancisco_data %>%
  select(NatWalkInd, CSA_Name_ABBREV, Ac_Unpr, P_WrkAge, P_WrkAge, Pct_A00, Pct_A01, Pct_A02p, R_LowWag)

plot_1 <- suppressWarnings(GGally::ggpairs(NY_SF_focus[, c(1, 3:7)],
  aes(color = NY_SF_focus$CSA_Name_ABBREV),
  title = "Scatterplot marix of NatWalkability Index Grouped by city",
  subtitle = "Using the first 4 potential predictors"))

## Registered S3 method overwritten by 'GGally':
##   method from
##     +.gg   ggplot2

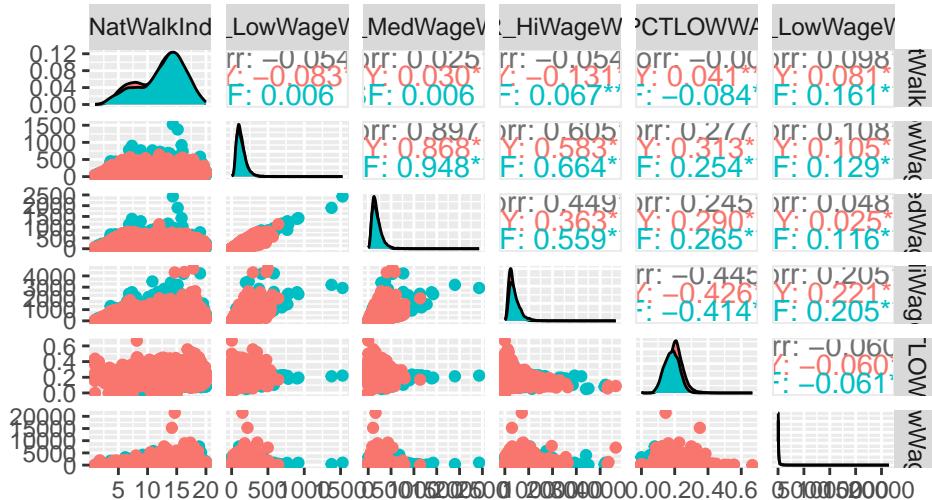
plot_1
```

Scatterplot marix of NatWalkability Index Grouped by city



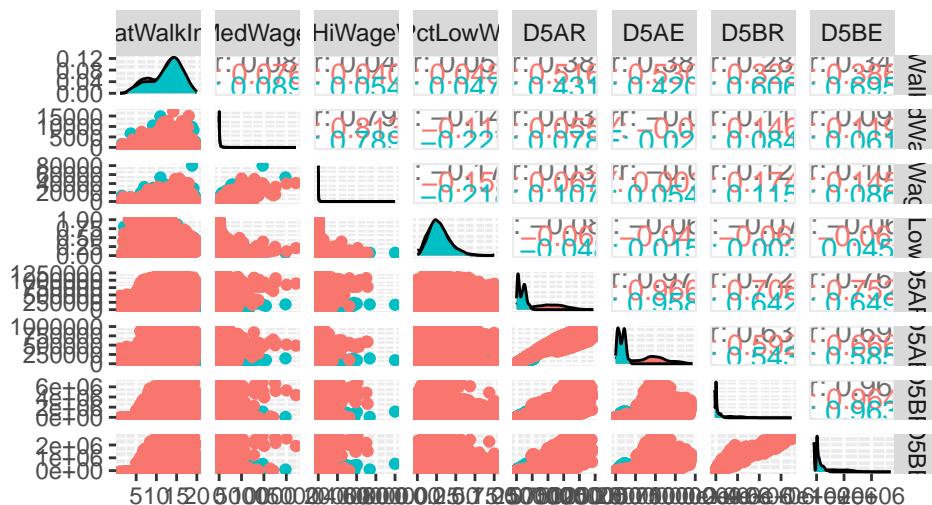
```
GGally::ggpairs(NY_SF_focus[, c(1, 8:12)],
  aes(color = NY_SF_focus$CSA_Name_ABBREV),
  title = "Scatterplot marix of NatWalkability Index Grouped by city",
  subtitle = "Using the next 5 potential predictors")
```

Scatterplot marix of NatWalkability Index Grouped by city



```
GGally::ggpairs(NY_SF_focus[, c(1, 13:19)],
  aes(color = NY_SF_focus$CSA_Name_ABBREV),
  title = "Scatterplot marix of NatWalkability Index Grouped by city",
  subtitle = "Using the following 5 potential predictors")
```

Scatterplot matrix of NatWalkability Index Grouped by c



```
GGally:::ggpairs(NY_SF_focus[, c(1, 20:24)],  
  aes(color = NY_SF_focus$CSA_Name_ABBREV),  
  title = "Scatterplot marix of NatWalkability Index Grouped by city",  
  subtitle = "Using the last 4 potential predictors")
```

Scatterplot marix of NatWalkability Index Grouped by city

