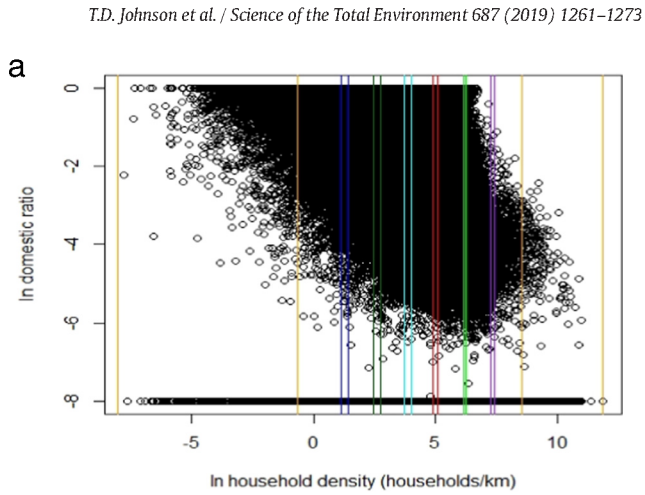
Recreating figure 2 (panel a) from Johnson et al., 2019



We recreated figure 2, panel (a) from Johnson et al. (2019) by using the original 1990 Census data for households and source of water. Below is a full walkthrough of that recreation, along with some implications as to how it relates to our proposed methods.

```
# Johnson et al use household density as opposed to housing unit density
## Load Household data from 1990 Census
hshlds <- read.csv(here("data/tables/nhgis0253_ds120_1990_blck_grp.csv"))%>% #File directory
  select(GISJOIN, EU0001) # Columns for join field and households
colnames(hshlds) <- c("GISJOIN","Households") # rename columns for easy use

# Pull in source of water data from 1990 Census
sow <- read.csv(here("data/tables/nhgis_ds123_1990_blck_grp.csv"))%>% # File directory
  select(GISJOIN,EX5001,EX5002,EX5003,EX5004) # Select columns for sow
colnames(sow) <- c("GISJOIN","Public","Drilled","Dug","Other") # rename columns for easy use

# Load the spatial data for density calculations
spatial <- sf::st_read(here("data/shapefiles/US_blck_grp_1990.shp"))%>%
  sf::st_drop_geometry() # drop the geometry to make file size smaller
```

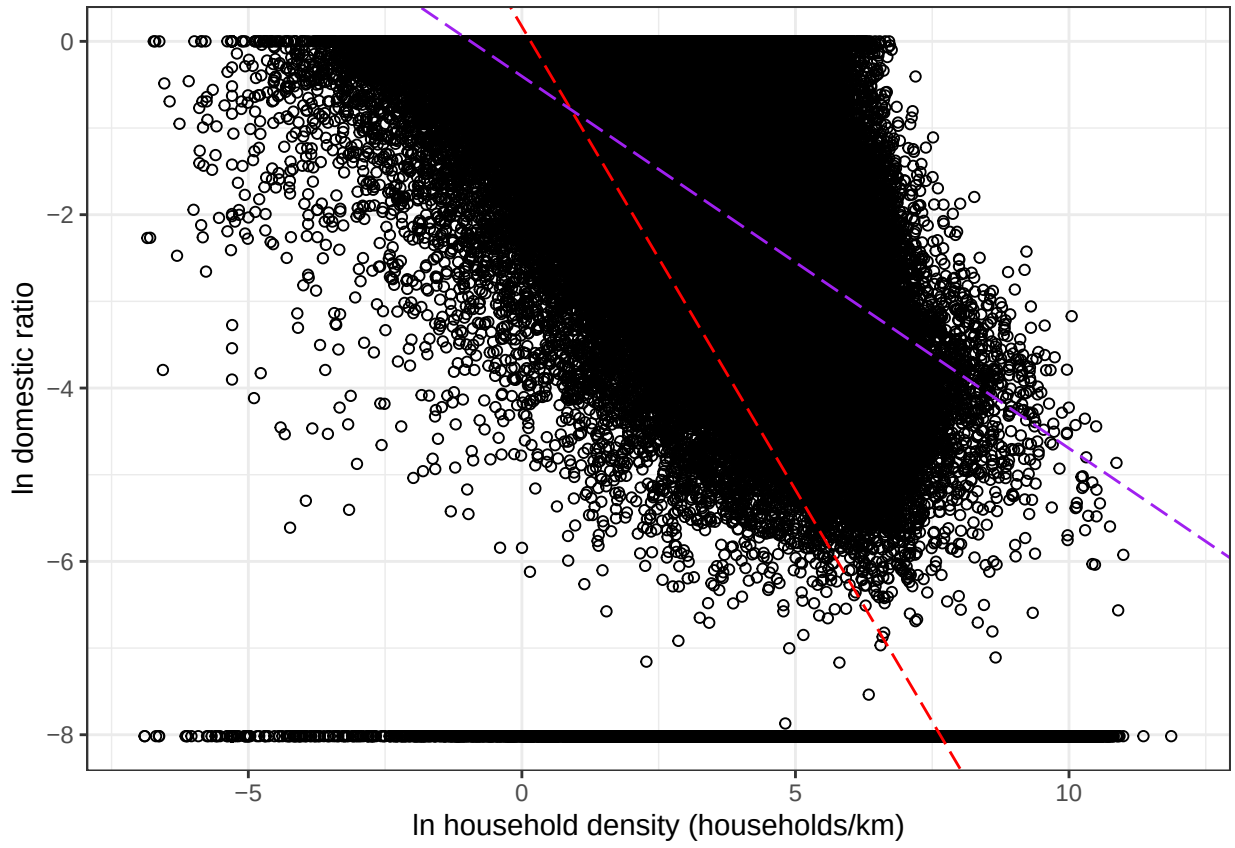
```
## Reading layer `US_blck_grp_1990' from data source `/proj/diegorilab/users/Andrew/DomesticWells/data/s
## Simple feature collection with 226388 features and 8 fields
## geometry type:  MULTIPOLYGON
## dimension:      XY
## bbox:           xmin: -7115713 ymin: -1337508 xmax: 2258225 ymax: 4591616
## proj4string:    +proj=aea +lat_1=29.5 +lat_2=45.5 +lat_0=37.5 +lon_0=-96 +x_0=0 +y_0=0 +datum=NAD83 +
```

```
# Combine everything and create ratios for plotting - replace zeroes with .0003
all <- spatial%>%
  select(GISJOIN, SHAPE_AREA)%>% # We only need the join field and the block group area
  left_join(hshlds)%>% # join household data
  left_join(sow)%>% # join source of water data
  mutate(hshld_density = Households / (SHAPE_AREA/1000000),   # calculate household density
         Domestic_Ratio = (Drilled+Dug) / (Public+Drilled+Dug+Other))%>% # Calculate the DR
  mutate(hshld_density = ifelse(hshld_density == 0, .005,hshld_density),  # renumber zeroes
         Domestic_Ratio = ifelse(Domestic_Ratio == 0, .00033,Domestic_Ratio)) # renumber zeroes
```

Replicated Plot

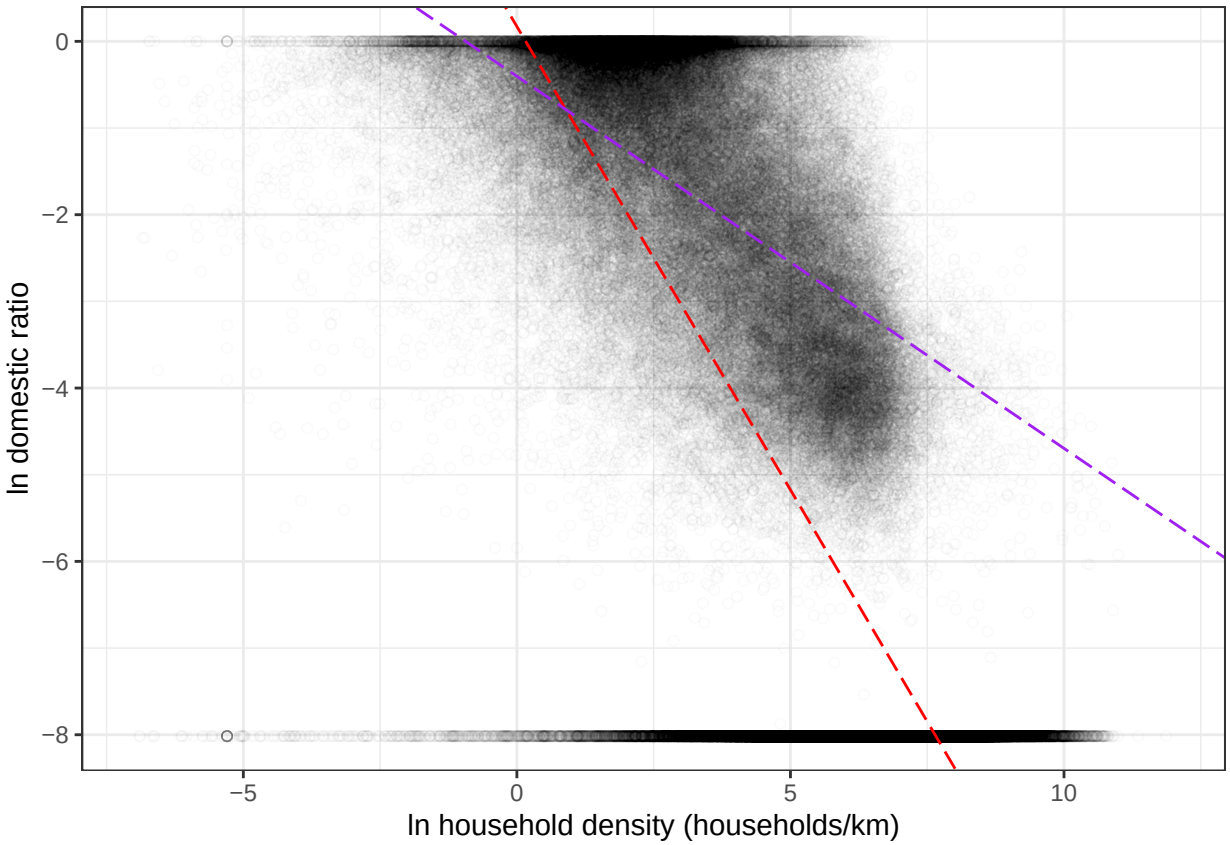
Here you can see the replicated plot relating the natural log of household density on the x-axis to the natural log of the domestic ratio on the y-axis. It should be noted that the domestic ratio is zero for 124,881 out of 226,388 block groups. These points can be observed along the bottom of the plot (Y = -8). The natural log of zero is infinity so it appears that Johnson et al have changed the zeroes to ~.00033 so that they are at the bottom of the plot. This is the value we used to replicate their plot, although the exact value used was not given in the paper. The r^2 value for this relationship is .057 when including block groups with a domestic ratio of zero and .037 when excluding these same block groups. We have added the regression lines to this plot for context (red includes zeroes as .0003 / purple removes all zeroes.)

```
# Red is .0003 instead of 0, purple is zeroes removed
ggplot(all)+ # Create plot
  geom_point(aes(x = log(hshld_density), y = log(Domestic_Ratio)), alpha = 1, shape = 1)+ # Add points
  geom_abline(slope = -1.069168, intercept = 0.171263, color = "red", linetype = "longdash")+ # add reg
  geom_abline(slope = -0.429318, intercept = -0.402770, color = "purple", linetype = "longdash")+ # add
  xlim(-7,12)+ # Limit the display to match Johnson et al
  labs(x = "ln household density (households/km)", # match labels
       y = "ln domestic ratio")+
  theme_bw() # match theme
```



By adding in some transparency to this plot, we can more clearly see the distribution of points.

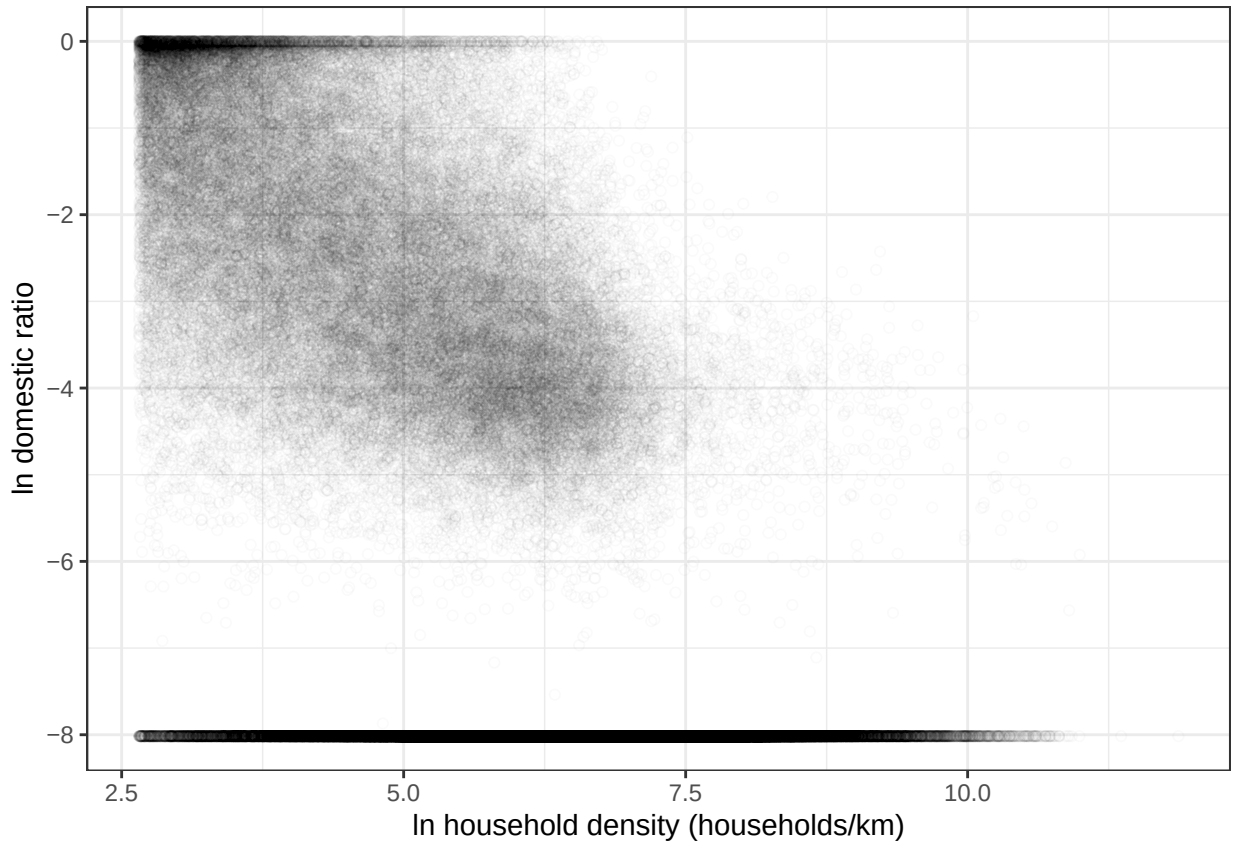
```
ggplot(all)+ # Same figure as above but with transparency
  geom_point(aes(x = log(hshld_density), y = log(Domestic_Ratio)), alpha = .02, shape = 1)+
  geom_abline(slope = -1.069168, intercept = 0.171263, color = "red", linetype = "longdash")+
  geom_abline(slope = -0.429318, intercept = -0.402770, color = "purple", linetype = "longdash")+
  xlim(-7,12)+
  labs(x = "ln household density (households/km)",
       y = "ln domestic ratio")+
  theme_bw()
```



What comes into focus is that there are essentially two data clusters here, the points along the bottom (Y =-8) represent 55% of the block groups in 1990, which is driving the relationship presented between domestic ratio and household density. Johnson et al set a threshold of 14.2 households per km^2 where block groups below that threshold maintain the DR from 1990 (or 2000). Therefore, the crux of the Johnson et al method is that in block groups, where household density is greater than 14.2 households per km^2 the Domestic ratio decreases as household density decreases. This would necessitate a strong relationship between household density and the domestic ratio. We can drill down on this part of the plot to take a deeper look.

```
# Filter so that we retain only block groups with >= 14.2 households per km^2
filt <- all%>%
  filter(log(hshld_density) >= 2.653)    # The natural log of 14.2 is 2.653

# Plot the subset
ggplot(filt)+
  geom_point(aes(x = log(hshld_density), y = log(Domestic_Ratio)), alpha = .02, shape = 1)+
  labs(x = "ln household density (households/km)",
       y = "ln domestic ratio")+
  theme_bw()
```



There are clearly two main clusters here (most likely Urban on the bottom and Suburban -> rural on top). We can look at the relationships both globally and we can bifurcate the data to take a closer look. The global relationship is shown below, with an r2 value of .032

Global Relationship of block groups with >= 14.2 households per km2

```
# run the global regression -- in R ln() is calculated by using the log() function with the default bas
reg <- lm(log(filt$Domestic_Ratio)~log(filt$hshld_density))
summary(reg) # Output the regression results
```

```
##
## Call:
## lm(formula = log(filt$Domestic_Ratio) ~ log(filt$hshld_density))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7799 -1.3443 -0.4237  1.4704  8.0261
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                0.977781   0.020585    47.5   <2e-16 ***
## log(filt$hshld_density) -1.211491   0.003353  -361.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 2.083 on 179957 degrees of freedom
##   (212 observations deleted due to missingness)
## Multiple R-squared:  0.4205, Adjusted R-squared:  0.4205
## F-statistic: 1.306e+05 on 1 and 179957 DF,  p-value: < 2.2e-16
```

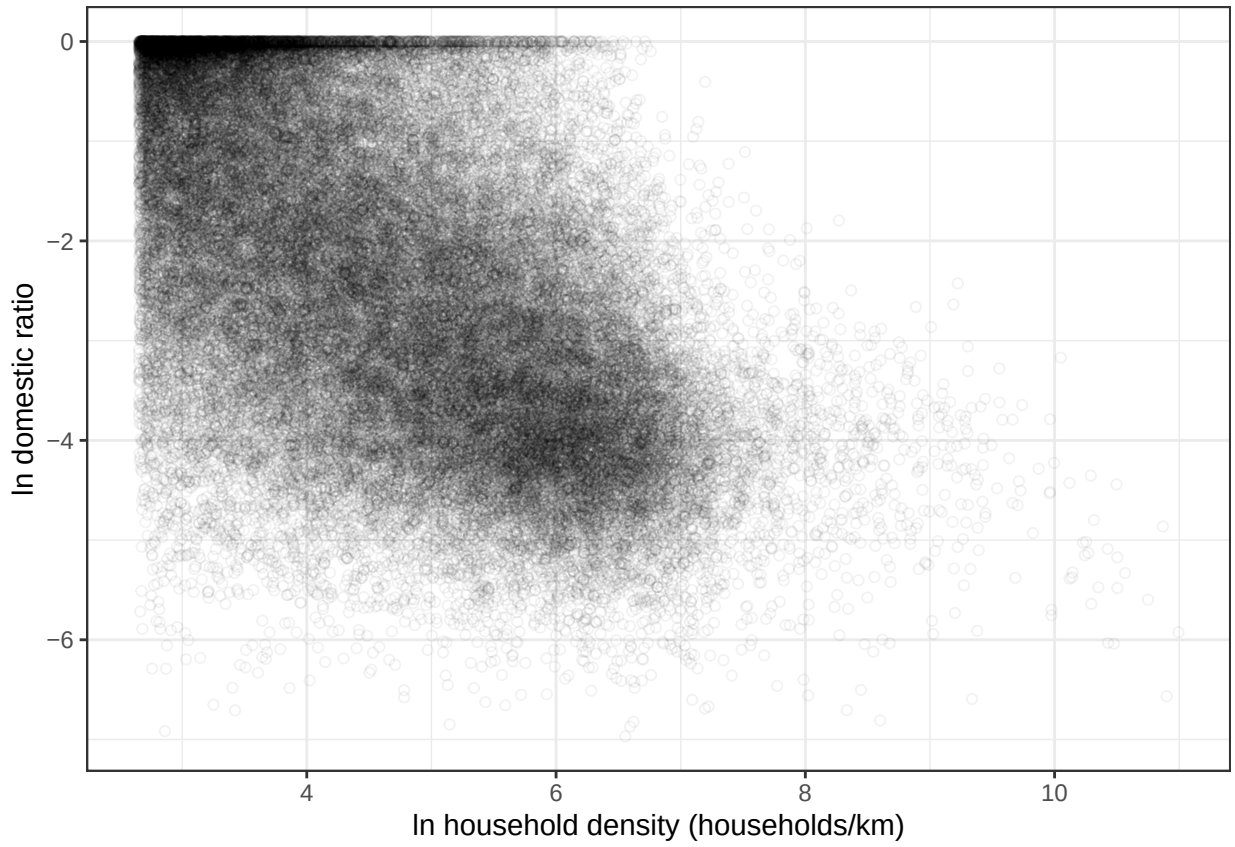
Bifurcating the data

Here, we bifurcate the data using the natural log of the domestic ratio >< -7

Greater than or equal to ln(Domestic Ratio) = -7:

```
filtGT7 <- filt%>% # filter so that we only keep rural block groups (ln(DR) > -7)
  filter(log(Domestic_Ratio)>=-7)

ggplot(filtGT7)+
  geom_point(aes(x = log(hshld_density), y = log(Domestic_Ratio)), alpha = .05, shape = 1)+
  labs(x = "ln household density (households/km)",
       y = "ln domestic ratio")+
  theme_bw()
```




```
summary(lm(log(filtGT7$Domestic_Ratio)~log(filtGT7$hshld_density))) # Output regression results
```

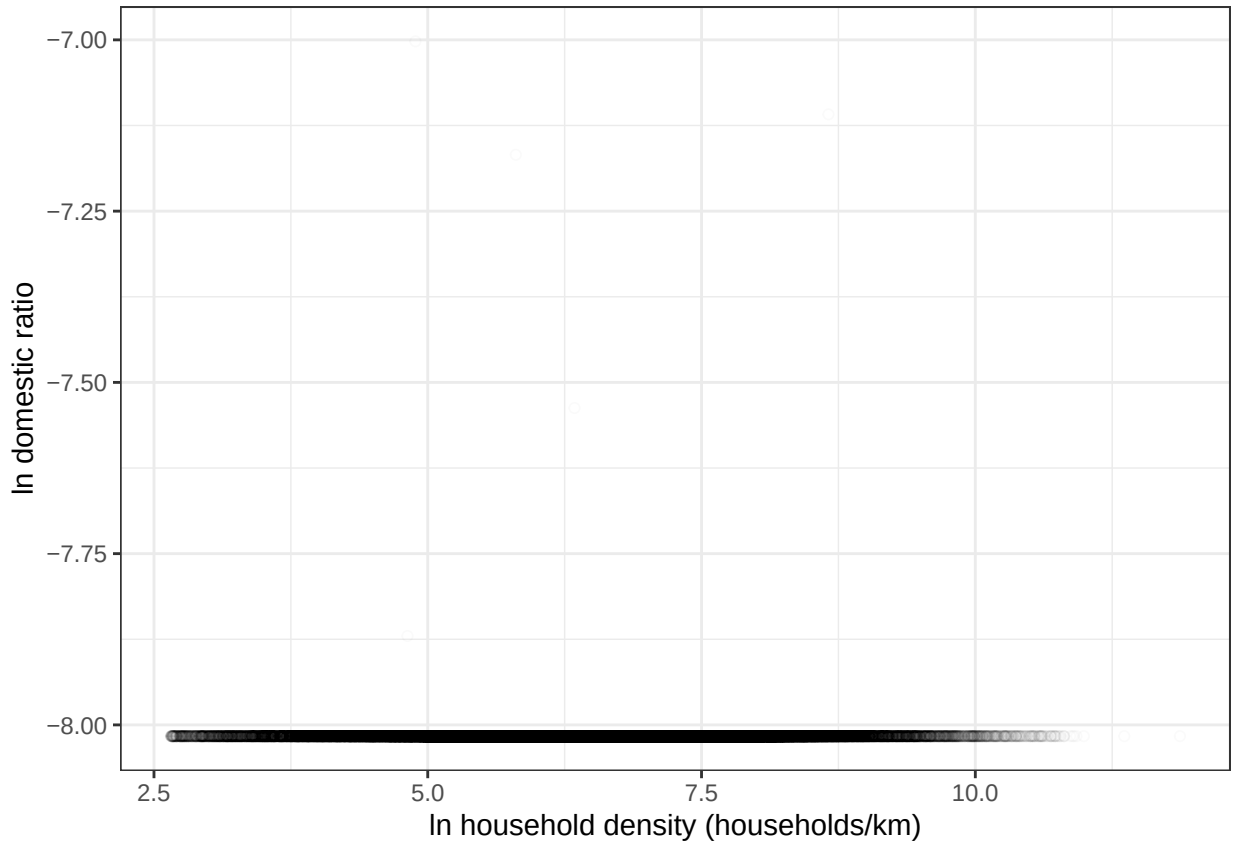
```
##
## Call:
```

```
## lm(formula = log(filtGT7$Domestic_Ratio) ~ log(filtGT7$hshld_density))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6579 -0.9129  0.0067  0.9958  3.7412
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  0.585527   0.020015   29.25   <2e-16 ***
## log(filtGT7$hshld_density) -0.644588   0.004109 -156.86   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.284 on 57122 degrees of freedom
## Multiple R-squared:  0.3011, Adjusted R-squared:  0.3011
## F-statistic: 2.461e+04 on 1 and 57122 DF,  p-value: < 2.2e-16
```

Less than ln(Domestic Ratio) = -7:

```
filtLT7 <- filt%>% # Filter so we only retain urban block groups (ln(DR) < -7)
  filter(log(Domestic_Ratio) < -7)

ggplot(filtLT7)+
  geom_point(aes(x = log(hshld_density), y = log(Domestic_Ratio)), alpha = .02, shape = 1)+
  labs(x = "ln household density (households/km)",
       y = "ln domestic ratio")+
  theme_bw()
```




```
summary(lm(log(filtLT7$Domestic_Ratio)~log(filtLT7$hshld_density))) # Output regression results
```

```
##
## Call:
## lm(formula = log(filtLT7$Domestic_Ratio) ~ log(filtLT7$hshld_density))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.00005 -0.00003 -0.00003 -0.00002  1.01423
##
## Coefficients:
##                               Estimate Std. Error    t value Pr(>|t|)
## (Intercept)                 -8.016e+00  8.139e-05 -98493.519   <2e-16 ***
## log(filtLT7$hshld_density) -5.026e-06  1.224e-05     -0.411    0.681
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.004795 on 122833 degrees of freedom
## Multiple R-squared:  1.372e-06,  Adjusted R-squared:  -6.769e-06
## F-statistic: 0.1686 on 1 and 122833 DF,  p-value: 0.6814
```

What this shows is that there is bias introduced from urban block groups where the domestic ratio is zero and household density is very high. This creates the appearance of a relationship between household density and the domestic ratio when viwed globally, however the relationship is essentially nonexistent when you isolate block groups above the Johnson threshold of 14.2 and bifurcate the data between urban and rural/suburban block groups.