Census EDA

Scott White

2019-04-18

```r
library(Hmisc)
library(tidyverse)
library(broom)
library(gridExtra)
library(here)
```

```r
cleaned <- read_csv(here("Final Project", "data_output", "cleaned.csv"),
                    col_types = cols(`Census Year` = col_factor()))
```

```r
summary(cleaned)
```

There are two variables that are missing 237 values. Upon inspection, those are the values for 2001, so we'll remove those variables for now.

```r
cleaned = cleaned %>% select(-`Income All Families Median`,
                             -`Income All Families Standard Error`)
```

```r
table(cleaned$`Boundary Type`)
```

```
##
##                  City        Community Area         Neighbourhood
##                     8                    24                   358
## Neighbourhood Cluster                  Ward
##                    46                    30
```

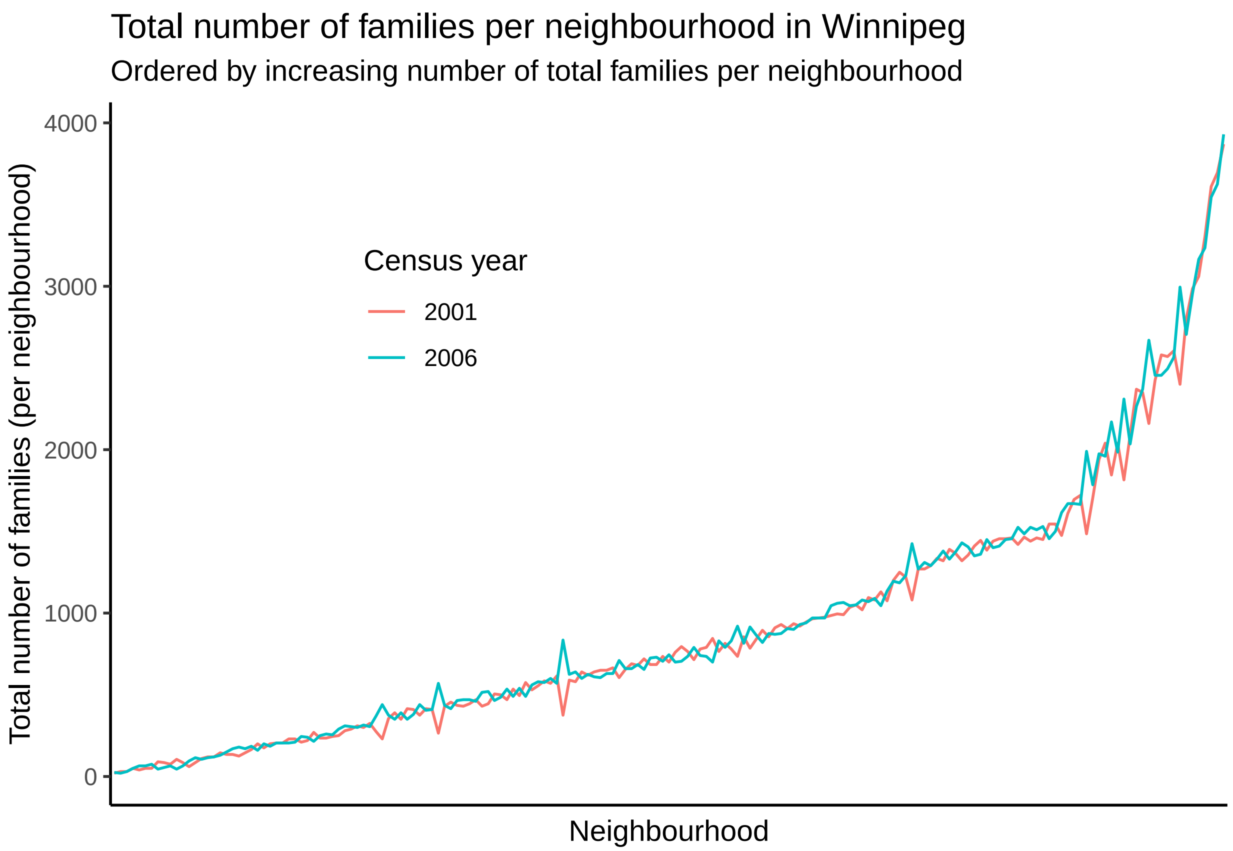
Lets begin by investigating the data just related to the Neighbourhood Boundary Type.

```r
neighbourhood <- cleaned %>% filter(`Boundary Type` == "Neighbourhood")
```

This first part of the analysis involves assessing if large families don't move as much as smaller families.

An intital check to see if there are any obvious population differences between years.

```r
plot_neighbourhood_sizetotalfamilies <-  neighbourhood %>% ggplot(aes(x = reorder(`Boundary Name`,
                                         `Size Total Families`),
                             group = `Census Year`)) +
  geom_line(aes(y = `Size Total Families`, color = `Census Year`)) +
  theme_classic() +
  theme(axis.text.x = element_blank(),
        axis.ticks.x = element_blank(),
        legend.position = c(0.3, 0.7)) +
  scale_color_discrete(name = "Census year") +
  xlab("Neighbourhood") +
  ylab("Total number of families (per neighbourhood)") +
  ggtitle("Total number of families per neighbourhood in Winnipeg",
          "Ordered by increasing number of total families per neighbourhood")
plot_neighbourhood_sizetotalfamilies
```



It doesn't appear that there is a huge change between 2001 and 2006. The graph does appear to grow in a linear fashion up to a point and then grow at a much faster rate. So it appears that an exponential model would characterize the number of families an area has. Below we plot the 2001 values versus the 2006 values to show that they do follow a fairly linear relationship.

Below we fit a polynomial regression to the data. Since the data seem to follow a fairly linear relation the model will be fit to the average number of families in an area. Also, since the x axis is technically not a quantitative variable, we will regress the number of families against the neighbourhood rank. In this case the higher the rank the larger the number of families in the area.

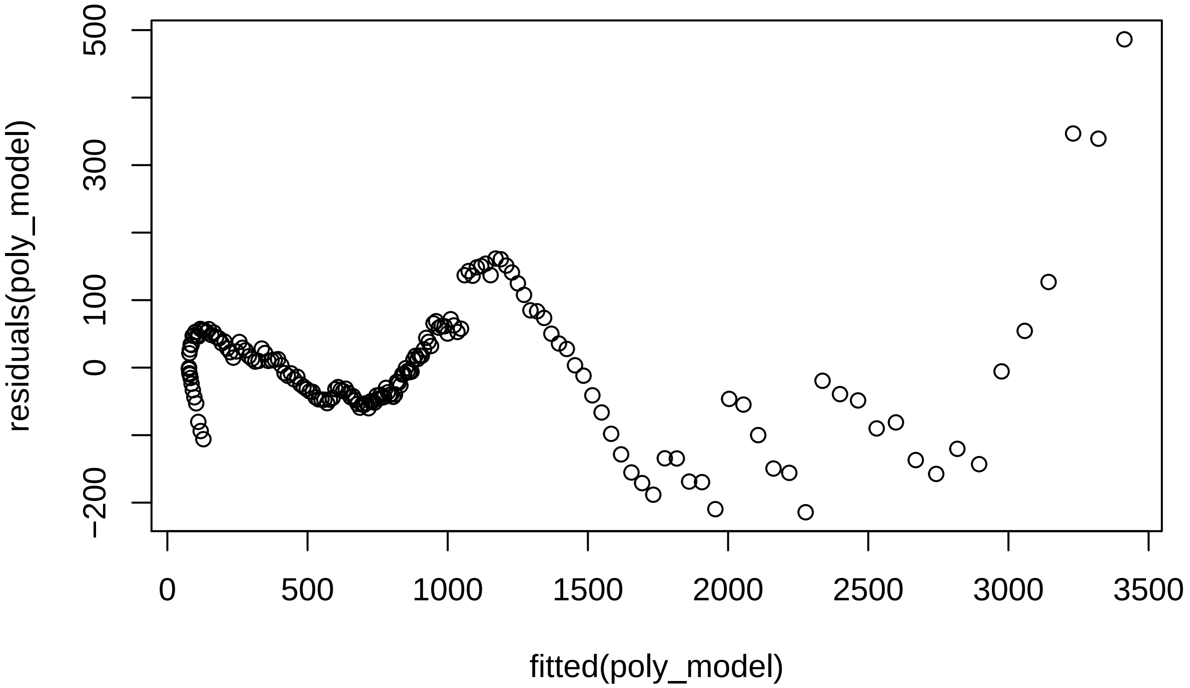
```
mean_rank <- neighbourhood %>% group_by(`Boundary Name`) %>%
  summarise(Mean = mean(`Size Total Families`)) %>%
  arrange(Mean) %>%
  mutate(Rank = seq(length(Mean)))
mean_rank
```

```
## # A tibble: 179 x 3
##    `Boundary Name`      Mean  Rank
##    <chr>               <dbl> <int>
##  1 South Point Douglas  22.5     1
##  2 Tissot               25       2
##  3 Holden               30       3
##  4 Polo Park            50       4
##  5 Logan-C.P.R.         52.5     5
##  6 West Wolseley        57.5     6
##  7 Exchange District    62.5     7
##  8 Airport              67.5     8
##  9 Cloutier Drive       70       9
## 10 Kensington           70      10
## # ... with 169 more rows
```

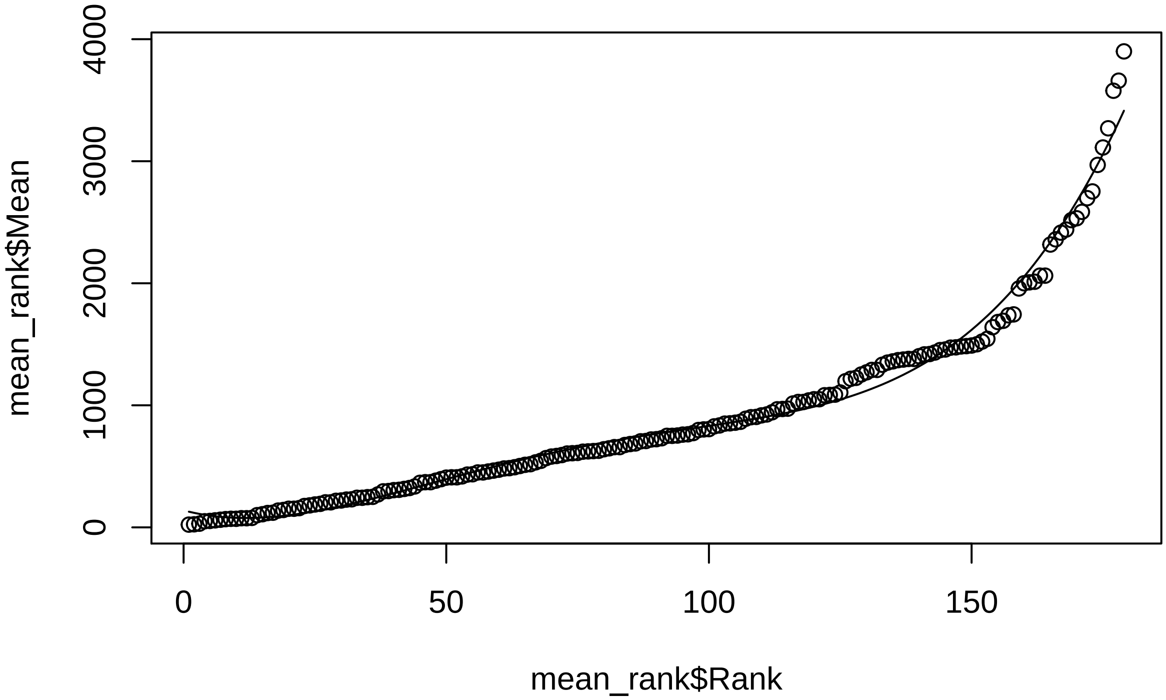
```
poly_model <- lm(Mean ~ poly(Rank, 4), data = mean_rank)
summary(poly_model)
```

```
## 
## Call:
## lm(formula = Mean ~ poly(Rank, 4), data = mean_rank)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -214.19  -44.49   -6.62   44.72  486.31 
## 
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)    
## (Intercept)     930.628      6.831  136.24   <2e-16 ***
## poly(Rank, 4)1 9557.349     91.388  104.58   <2e-16 ***
## poly(Rank, 4)2 3342.569     91.388   36.58   <2e-16 ***
## poly(Rank, 4)3 2155.512     91.388   23.59   <2e-16 ***
## poly(Rank, 4)4 1372.810     91.388   15.02   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 91.39 on 174 degrees of freedom
## Multiple R-squared:  0.9868, Adjusted R-squared:  0.9865 
## F-statistic:  3264 on 4 and 174 DF,  p-value: < 2.2e-16
```

```
plot(fitted(poly_model), residuals(poly_model))
```

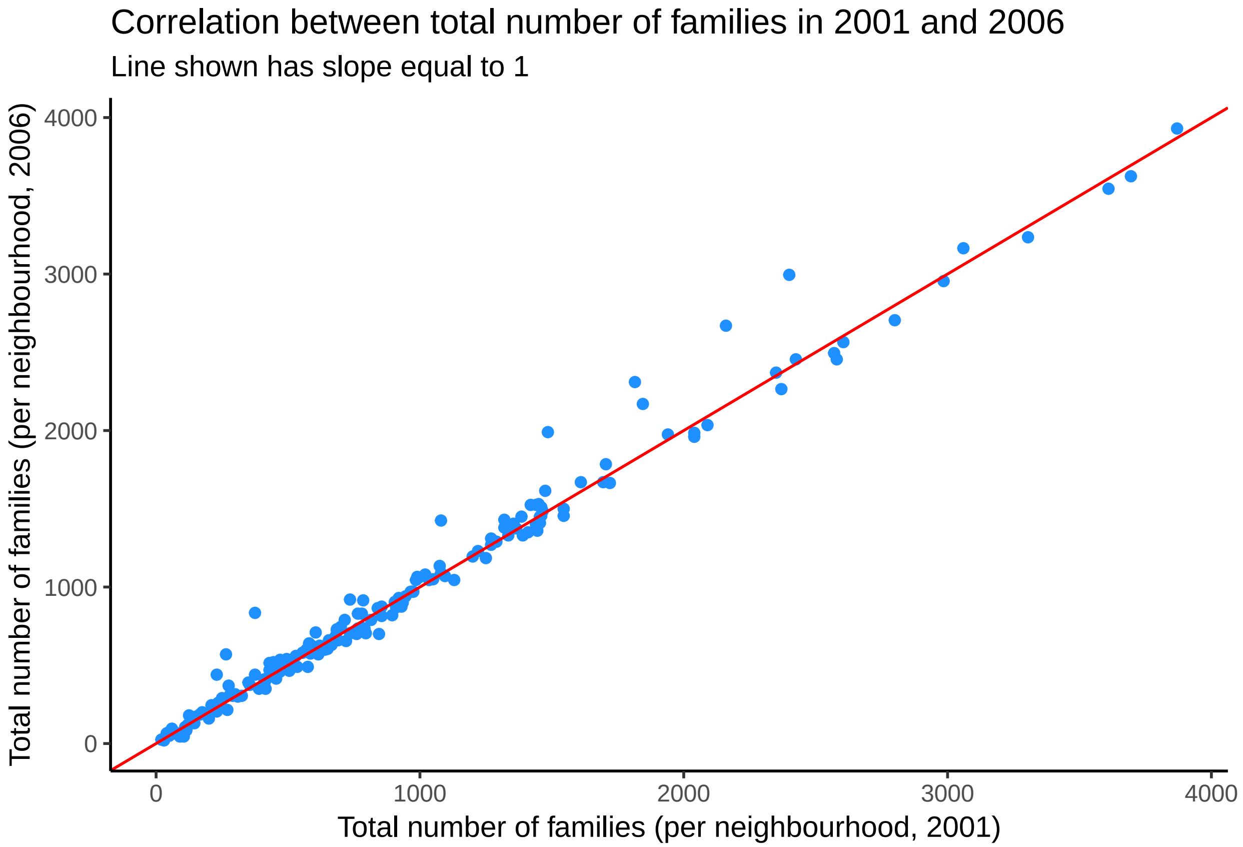


```
plot(mean_rank$Rank, mean_rank$Mean)
lines(mean_rank$Rank, fitted(poly_model))
```



Though the quaratic plot fits the data fairly well, the residuals are all over the place and so this probably isn't a great technique to use to generalize the data.

```
neighbourhood_total_family_size_counts <- neighbourhood %>% group_by(`Boundary Name`) %>%
  summarise(`2001 Count` = first(`Size Total Families`),
            `2006 Count` = last(`Size Total Families`))

plot_yeartoyear_linear_relationship <- neighbourhood_total_family_size_counts%>%
  ggplot(aes(x = `2001 Count`, y = `2006 Count`)) +
  geom_point(color = "Dodger blue") +
  geom_abline(slope = 1, color = "red") +
  theme_classic() +
  xlab("Total number of families (per neighbourhood, 2001)") +
  ylab("Total number of families (per neighbourhood, 2006)") +
  ggtitle("Correlation between total number of families in 2001 and 2006",
          "Line shown has slope equal to 1")
plot_yeartoyear_linear_relationship
```



There obviously isn't a perfect relationship, though a linear pattern is fairly consistent except for a few points that we can investigate. The neighbourhoods that are above the line have higher 2006 family sizes than the 2001 census counts, and vice versa. The following pearson correlation test supports the graphic further indicating a very strong linear relationship.

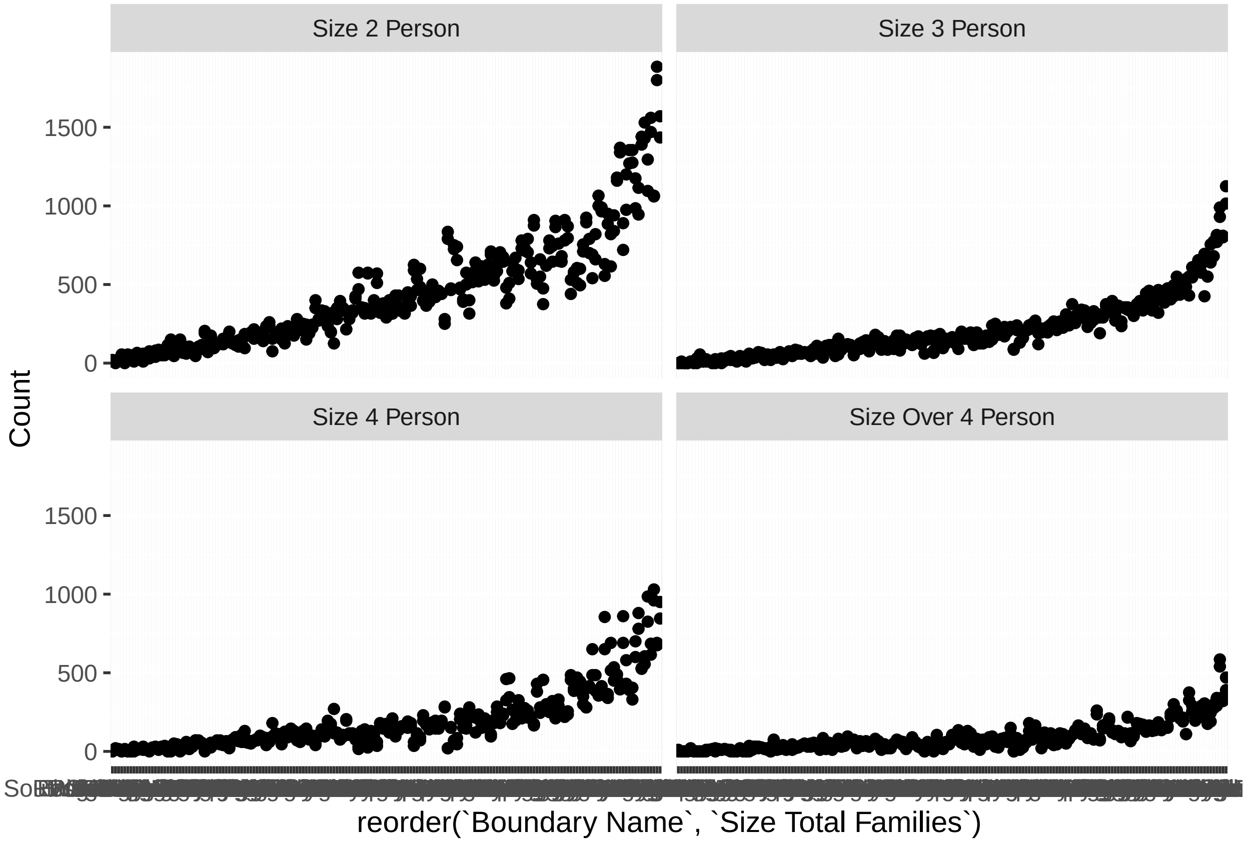
```
cor.test(neighbourhood_total_family_size_counts$`2001 Count`,
         neighbourhood_total_family_size_counts$`2006 Count`)
```

```
##
##  Pearson's product-moment correlation
##
## data:  neighbourhood_total_family_size_counts$`2001 Count` and neighbourhood_total_family_size_counts
## t = 97.439, df = 177, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9876663 0.9931507
## sample estimates:
##      cor
## 0.990807
```

TK: Test if the ratio is close to 1.

Lets check if there is a strong correlation between the total size of an areas family and the different sizes of families.

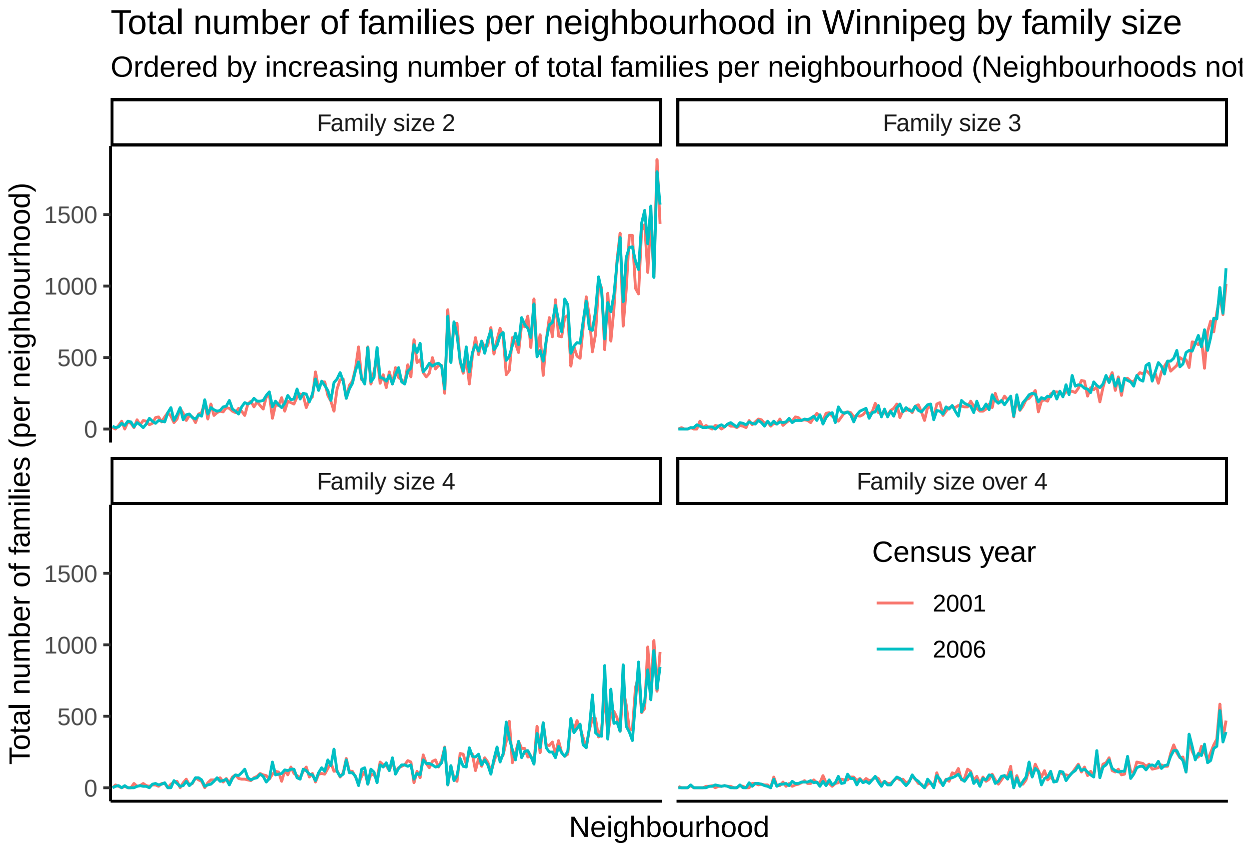
```
neighbourhood %>%
  select(`Boundary Name`, `Size 2 Person`:`Size Total Families`) %>%
  gather(key = "Family Size", value = "Count", `Size 2 Person`:`Size Over 4 Person`) %>%
  ggplot(aes(x = reorder(`Boundary Name`,
                         `Size Total Families`))) +
  geom_point(aes(y = Count)) +
    facet_wrap(.~`Family Size`)
```



Hmm. Those are some interesting patterns. It's clear that as the total number of families in an area increase so does the number of different size families. However, not all family sizes grow at the same rate. It appears that in general, as the number of families grow the smaller size families grow more quickly then the larger families. (Though the graph above includes data for both years)

Is there a pattern over years? Can I plot the difference between the two years and see how that difference changes?

```
total_facets = as_labeller(c("Size 2 Person" = "Family size 2", "Size 3 Person" = "Family size 3", "Siz
plot_facet_familysize_count <-  neighbourhood %>%
  select(`Census Year`,
         `Boundary Name`,
         `Size 2 Person`:`Size Total Families`) %>%
  gather(key = "Family Size", value = "Count",
         `Size 2 Person`:`Size Over 4 Person`) %>%
  ggplot(aes(x = reorder(`Boundary Name`,
                         `Size Total Families`),
             group = `Census Year`)) +
  geom_line(aes(y = Count, color = `Census Year`)) +
    facet_wrap(.~`Family Size`, labeller = total_facets) +
  theme_classic() +
  theme(axis.text.x = element_blank(),
        axis.ticks.x = element_blank(),
        legend.position = c(0.755, 0.3)) +
  xlab("Neighbourhood") +
  ylab("Total number of families (per neighbourhood)") +
  ggtitle("Total number of families per neighbourhood in Winnipeg by family size",
          "Ordered by increasing number of total families per neighbourhood (Neighbourhoods not listed)
  scale_color_discrete(name = "Census year")

plot_facet_familysize_count
```



From the graph above it doesn't appear that there is a lot of change In the family sizes for most areas. However, we can plot the differences and order the differences based on the number of family sizes to see if there really is little difference in the family size counts for each area. We will order the counts based on the average family size of both years combined size since the yearly counts generally follow a linear relationship with slope close to 1.

```
family_size_differences <- neighbourhood %>%
  select(`Boundary Name`, `Size 2 Person`:`Size Total Families`) %>%
  group_by(`Boundary Name`) %>%
  summarise(`Size 2 Diff`  = last(`Size 2 Person`) - first(`Size 2 Person`),
            `Size 3 Diff`  = last(`Size 3 Person`) - first(`Size 3 Person`),
            `Size 4 Diff`  = last(`Size 4 Person`) - first(`Size 4 Person`),
            `Size Over 4 Diff`  = last(`Size Over 4 Person`) - first(`Size Over 4 Person`),
            `Size Total Fam Avg`  = mean(`Size Total Families`))
family_size_differences
```

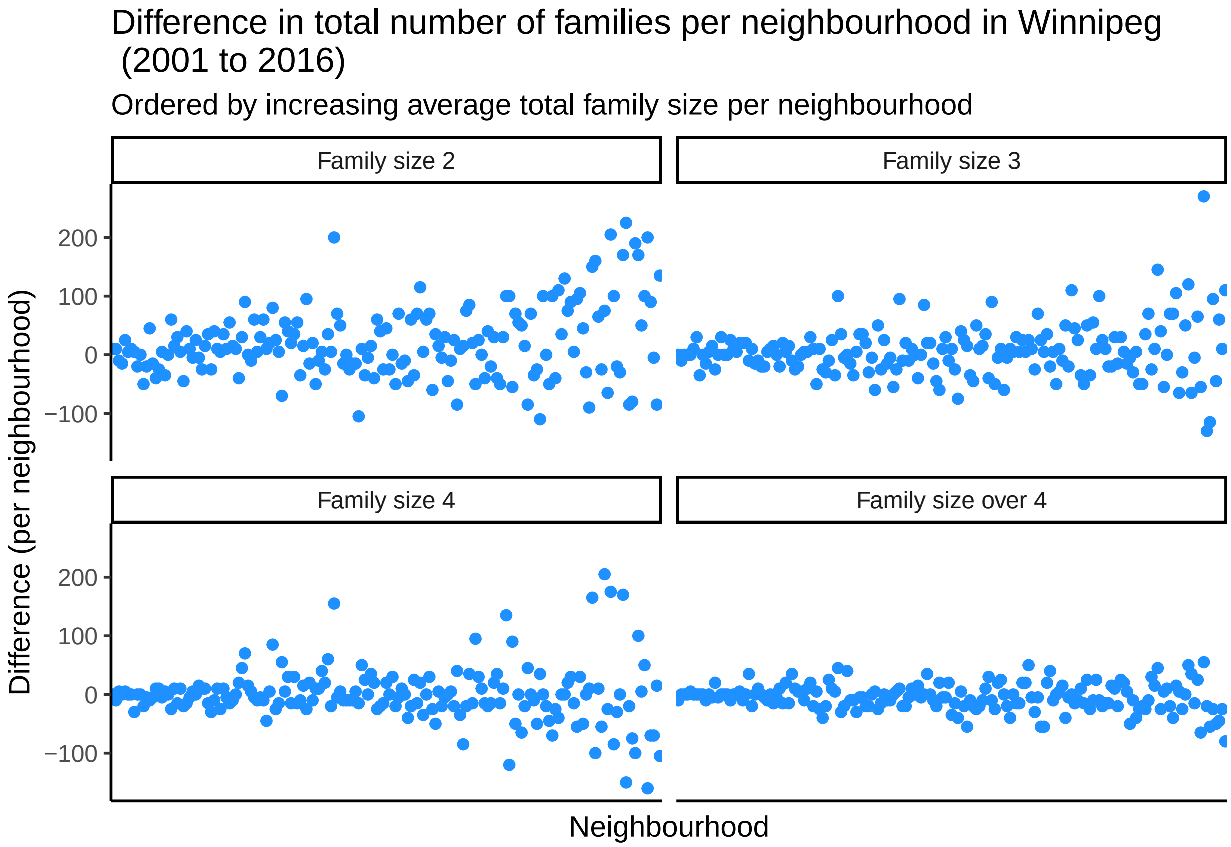
```
## # A tibble: 179 x 6
##    `Boundary Name` `Size 2 Diff` `Size 3 Diff` `Size 4 Diff`
##    <chr>                   <dbl>         <dbl>         <dbl>
##  1 Agassiz                   -35            25             5
##  2 Airport                     5           -35           -30
##  3 Alpine Place              -35            10            25
##  4 Amber Trails               80           100            85
##  5 Archwood                  -25             5            10
##  6 Armstrong Point           -40            30            10
##  7 Beaumont                  -25           -10           -15
##  8 Betsworth                 100            45          -120
##  9 Birchwood                 -50           -25            10
## 10 Booth                     -40            30           -25
## # ... with 169 more rows, and 2 more variables: `Size Over 4 Diff` <dbl>,
## #   `Size Total Fam Avg` <dbl>
```

The above values are calculated by taking the 2006 value and subtracting the 2001 value. Below I show the values for Agassiz to show that this is in fact what the calculations show.

```
neighbourhood %>% select(`Boundary Name`, `Size 2 Person`) %>%
  filter(`Boundary Name` == "Agassiz")
```

```
## # A tibble: 2 x 2
##   `Boundary Name` `Size 2 Person`
##   <chr>                     <dbl>
## 1 Agassiz                      85
## 2 Agassiz                      50
```

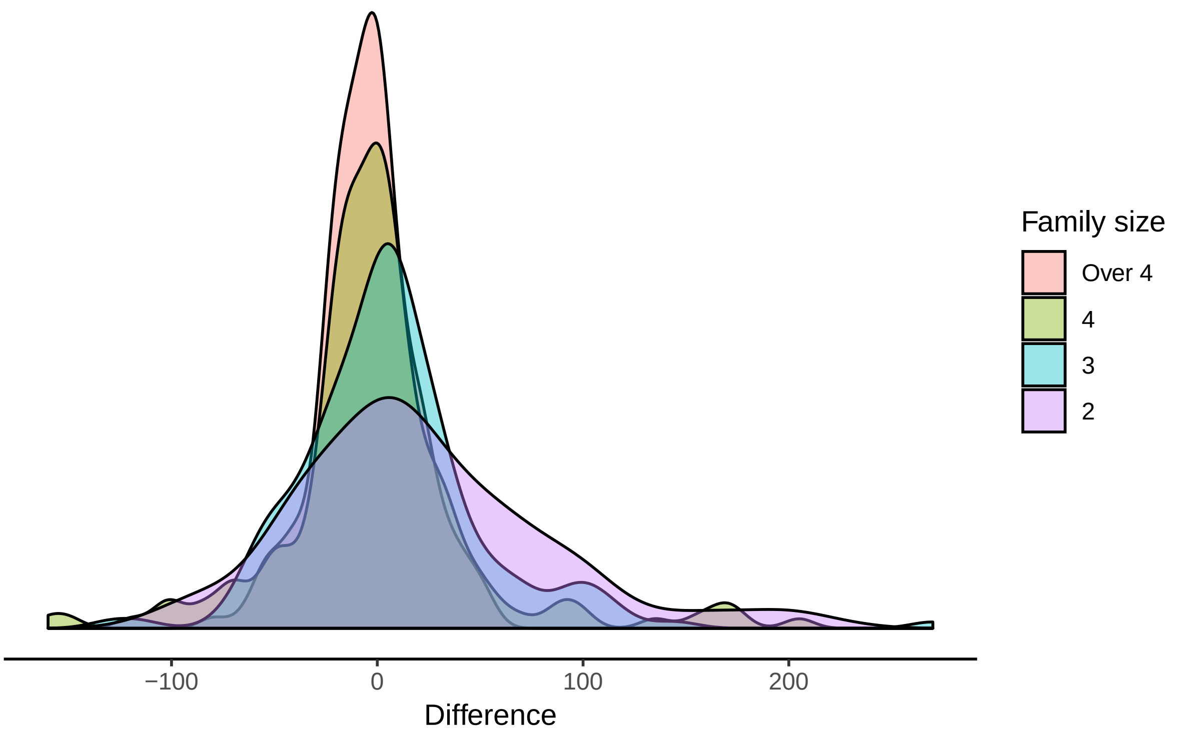
```
difference_facets = as_labeller(c("Size 2 Diff" = "Family size 2", "Size 3 Diff" = "Family size 3", "Si

plot_facet_familysize_differences <- family_size_differences %>%
  gather(key = "Family Size", value = "Difference",
         `Size 2 Diff`:`Size Over 4 Diff`) %>%
  ggplot(aes(x = reorder(`Boundary Name`,
                         `Size Total Fam Avg`))) +
  geom_point(aes(y = Difference), color = "Dodger blue") +
  theme_classic() +
  theme(axis.text.x = element_blank(),
        axis.ticks.x = element_blank()) +
  xlab("Neighbourhood") +
  ylab("Difference (per neighbourhood)") +
  ggtitle("Difference in total number of families per neighbourhood in Winnipeg \n (2001 to 2016)",
          "Ordered by increasing average total family size per neighbourhood") +
    facet_wrap(`Family Size`~., labeller = difference_facets)
plot_facet_familysize_differences
```



The above chart is fairly interesting.

The distribution of the differences appears to be normally distributed, and so we can overly density plots of the four family size categories to compare them. Below we do so and it appears as if the supposition may be correct. That is, the larger family sizes have smaller differences between years.

```
plot_density_differences <-  family_size_differences %>%
  gather(key = "Family Size", value = "Difference",
         `Size 2 Diff`:`Size Over 4 Diff`) %>%
  ggplot(aes(x = Difference, fill = fct_rev(`Family Size`))) +
  geom_density(alpha = 0.4) +
  theme_classic() +
  theme(axis.text.y = element_blank(),
        axis.ticks.y = element_blank(),
        axis.line.y = element_blank()) +
  xlab("Difference") +
  ylab("") +
  scale_fill_discrete(name = "Family size",
                      labels = c("Over 4", "4", "3", "2")) +
  ggtitle("Distributions of differences in total number of families by family size in Winnipeg \n (2001
plot_density_differences
```



Distributions of differences in total number of families by family size in Winnipe
(2001 to 2016)

We will apply the Shapiro-Wilks test for normality to check if the four different family size difference categories follow a normal distribution. We can easily do this by combining some functionality from both the broom and purrr packages.

```
data_shapiro_wilks <- family_size_differences %>%
  select(-`Boundary Name`, -`Size Total Fam Avg`) %>%
  gather(key = "FamSize", value = "Difference", `Size 2 Diff`:`Size Over 4 Diff`) %>%
  group_by(FamSize) %>%
  nest(-FamSize) %>%
  mutate(test = map(data, function(x) shapiro.test(x$Difference)   ),
         tidied = map(test, tidy)) %>%
  unnest(tidied, .drop = TRUE)
data_shapiro_wilks
```

```
## # A tibble: 4 x 4
##   FamSize          statistic  p.value method
##   <chr>                <dbl>    <dbl> <chr>
## 1 Size 2 Diff          0.958 3.73e- 5 Shapiro-Wilk normality test
## 2 Size 3 Diff          0.913 8.31e- 9 Shapiro-Wilk normality test
## 3 Size 4 Diff          0.870 2.59e-11 Shapiro-Wilk normality test
## 4 Size Over 4 Diff     0.981 1.77e- 2 Shapiro-Wilk normality test
```

```
write_csv(data_shapiro_wilks,
          path = here("Final Project", "data_output", "data_shapiro_wilks.csv"))
```

From the tests above we can see that the differences for the different family sizes between 2001 and 2006 are approximately normally distributed. The table below appears to send some support that the variance decreases as the family size increases. Families of size 2 have a much larger variance then the other family sizes, and family sizes greater than 4 have a significantly smaller variance of differences. The differences for families of sizes 3 and 4 don't appear to be fairly similar though, as shown by their close standard deviation.

```
data_difference_variances <- family_size_differences %>%
  select(-`Boundary Name`, -`Size Total Fam Avg`) %>%
  gather(key = "FamSize", value = "Difference", `Size 2 Diff`:`Size Over 4 Diff`) %>%
  group_by(FamSize) %>%
  summarise(Variance = var(Difference),
            StdDev = sd(Difference))
data_difference_variances
```

```
## # A tibble: 4 x 3
##   FamSize          Variance StdDev
##   <chr>               <dbl>  <dbl>
## 1 Size 2 Diff         3818.   61.8
## 2 Size 3 Diff         2098.   45.8
## 3 Size 4 Diff         2379.   48.8
## 4 Size Over 4 Diff     510.   22.6
```

```
write_csv(data_difference_variances,
          path = here("Final Project", "data_output", "data_difference_variances.csv"))
```

Bartlett's test can be used here to further support the non-homogeneity in variance.

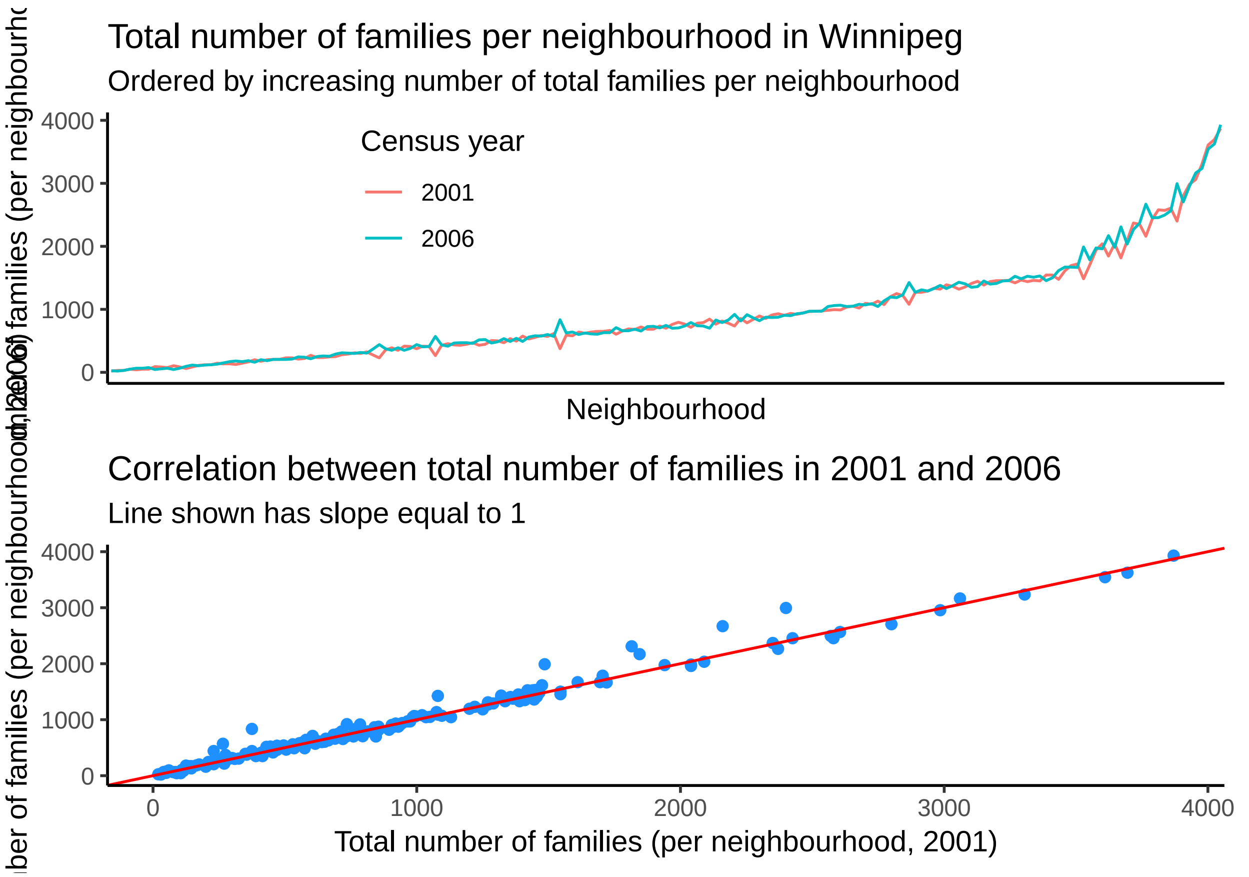
```
family_size_differences %>%
  select(-`Boundary Name`, -`Size Total Fam Avg`) %>%
  gather(key = "FamSize", value = "Difference", `Size 2 Diff`:`Size Over 4 Diff`) %>%
  bartlett.test(Difference ~ FamSize, data = .)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  Difference by FamSize
## Bartlett's K-squared = 156.6, df = 3, p-value < 2.2e-16
```

So the conclusions we can draw thus far is that though there is no significant difference in the different neighbourhoods family size there are differences in the rate of increase of the different family sizes as the area population increases. Meaning that in Winnipeg as the neighbourhood size grows it tends to grow greater than linear.

Below here I do the final combining of the plots for the final report.

```
plot_combined_sizetotalfamilies <- grid.arrange(plot_neighbourhood_sizetotalfamilies,
             plot_yeartoyear_linear_relationship,
             nrow = 2)
```

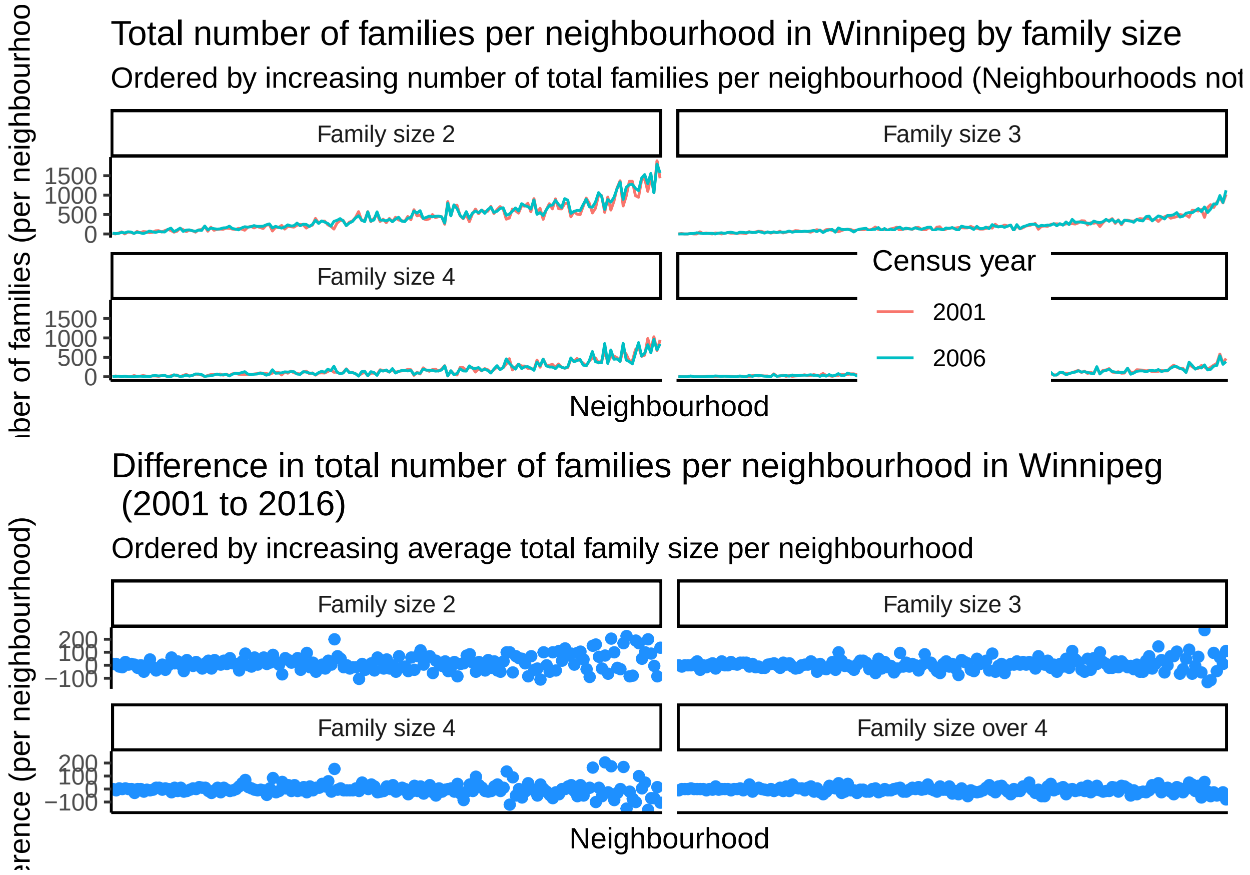


```
plot_combined_sizetotalfamilies
```

```
## TableGrob (2 x 1) "arrange": 2 grobs
##   z     cells    name           grob
## 1 1 (1-1,1-1) arrange gtable[layout]
## 2 2 (2-2,1-1) arrange gtable[layout]
```

```
ggsave(filename = here("Final Project", "figures", "plot_combined_sizetotalfamilies.png"),
       plot = plot_combined_sizetotalfamilies, width = 6, height = 10, units = "in")
```

```
plot_combined_differences <-  grid.arrange(plot_facet_familysize_count,
             plot_facet_familysize_differences,
             nrow = 2)
```

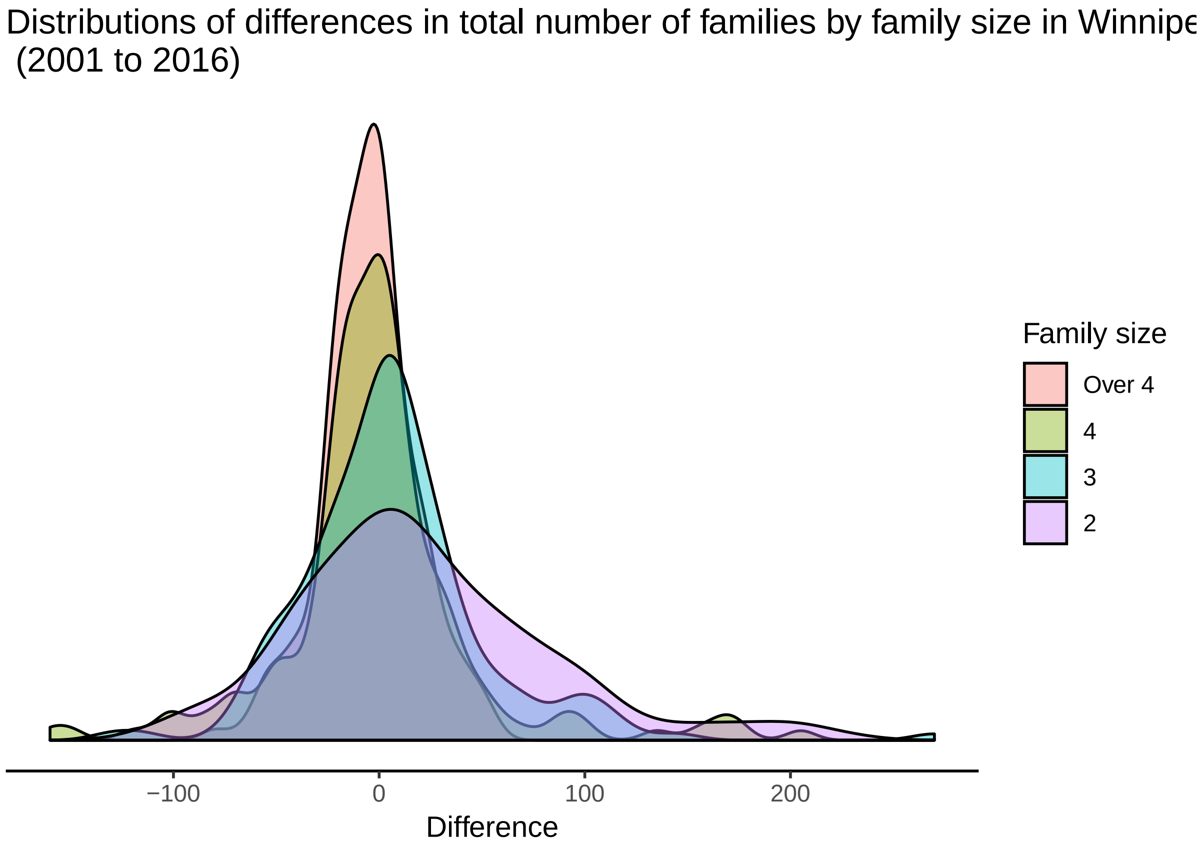


```
plot_combined_differences
```

```
## TableGrob (2 x 1) "arrange": 2 grobs
##   z     cells    name           grob
## 1 1 (1-1,1-1) arrange gtable[layout]
## 2 2 (2-2,1-1) arrange gtable[layout]
```

```
ggsave(filename = here("Final Project", "figures", "plot_combined_differences.png"),
       plot = plot_combined_differences,
       width = 6, height = 10, units = "in")
```

```
plot_density_differences
```



```
ggsave(filename = here("Final Project", "figures", "plot_density_differences.png"),
       plot = plot_density_differences)
```

```
## Saving 6.5 x 4.5 in image
```