Does Family Size Affect Mobility?

STAT 7350 - Final Project

Scott White

April 19, 2019

Introduction

Winnipeg contains a wide range of types of families, with one of the factors being family size. Some families are small, and others are very large. It might be simple to say that a families income affects their ability to move between different areas of a city, or even move out of the city. Family size is another factor that may influence this ability, and it is also more difficult to say if there is a relationship. The data to assess this is the City of Winnipeg census data for the years 2001 and 2006. So we will assess whether family size has a significant affect on a families mobility in this report.

Methods

The data set being analysed in this report is census data for the City of Winnipeg for 2001 and 2006 as it relates to income and family composition variables. The variables are counted for the different neighbourhoods of the city, as well as major sections of the city itself. We will be restricting our attention to the sections labelled neighbourhoods so that we can draw conclusions about the more distinct parts of the city. The data was taken from the City of Winnipeg Open Data Catalogue, and can be found here: https://data.winnipeg.ca/Census/Census-Census-Families/67cj-gsk9.

Pearson's correlation test was used to compare the counts for different areas for the years 2001 and 2006. This along with a visual inspection was used to check if the years had a linear relationship, as well to check if the diffence between the two years was not statistically different from zero. This approach was used to check if there was a significant change in the total number of families in a neighbourhood in 2001 and 2006.

We then look at the total number of families split into four categories: families of size two, three, four, or more than four. We visually check if the rate of family size increase is the same for all family size categories. Meaning, when we order the neighbourhoods by increasing order of the total number of families do the four categories increase at equal rates.

We then take a look at the change in total number of families of different sizes from 2001 to 2006 to compare the variation to give us an idea which types of families may be more consistent for particular areas. In other words, is there a greater variation in the change for families of size two than there are for family sizes over four? This will help us assess which family sizes are more stable in general for neighbourhoods, which will tell us if we should expect great swings in the number of those types of families. To answer this question we will use the Shapiro-Wilks test to first check the normality of these differences, and then use Bartlett's test to verify if there are differences in variances.

Results

Judging by the figure below, we can see that there does not appear to be a significant difference between the years 2001 and 2006. This is fairly interesting because it shows there is a consistency in the ratio between the two years independent of the magnitude of the neighbourhoods in question. Often in an analysis there is a greater variance for larger quantities and less so for smaller quantities, but that does not appear to be the case for the total number of families in the different neighbourhoods within Winnipeg. The second plot within Figure 1 shows that there is clustering around the line with slope 1, indicating a near 1:1 ratio between

the two census years. The Pearson correlation coefficient estimated a linear correlation of 99.2% between the two years.

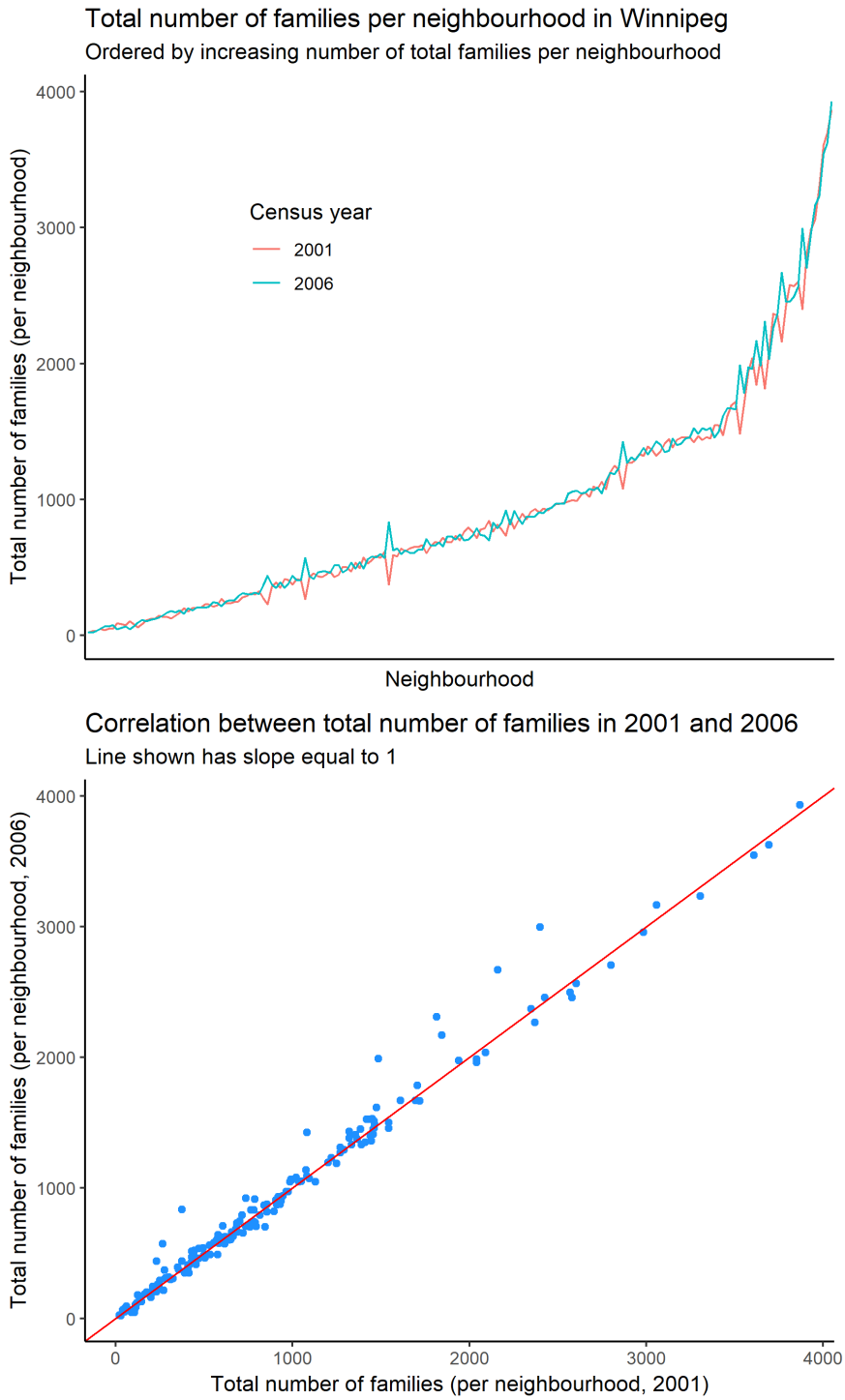

Figure 1: Top: Total number of families by neighbourhood ordered by increasing order of tital number of families in the individual neighbourhood. Bottom: Comparing the total number of families in each neighbourhood in both years 2001 and 2006, we should a strong linear correlation between them.

We now have some results when comparing neighbourhoods when we aggregate over the four different family size categories. Below we take a look at the differences when we separate the neighbourhoods by these different categories.

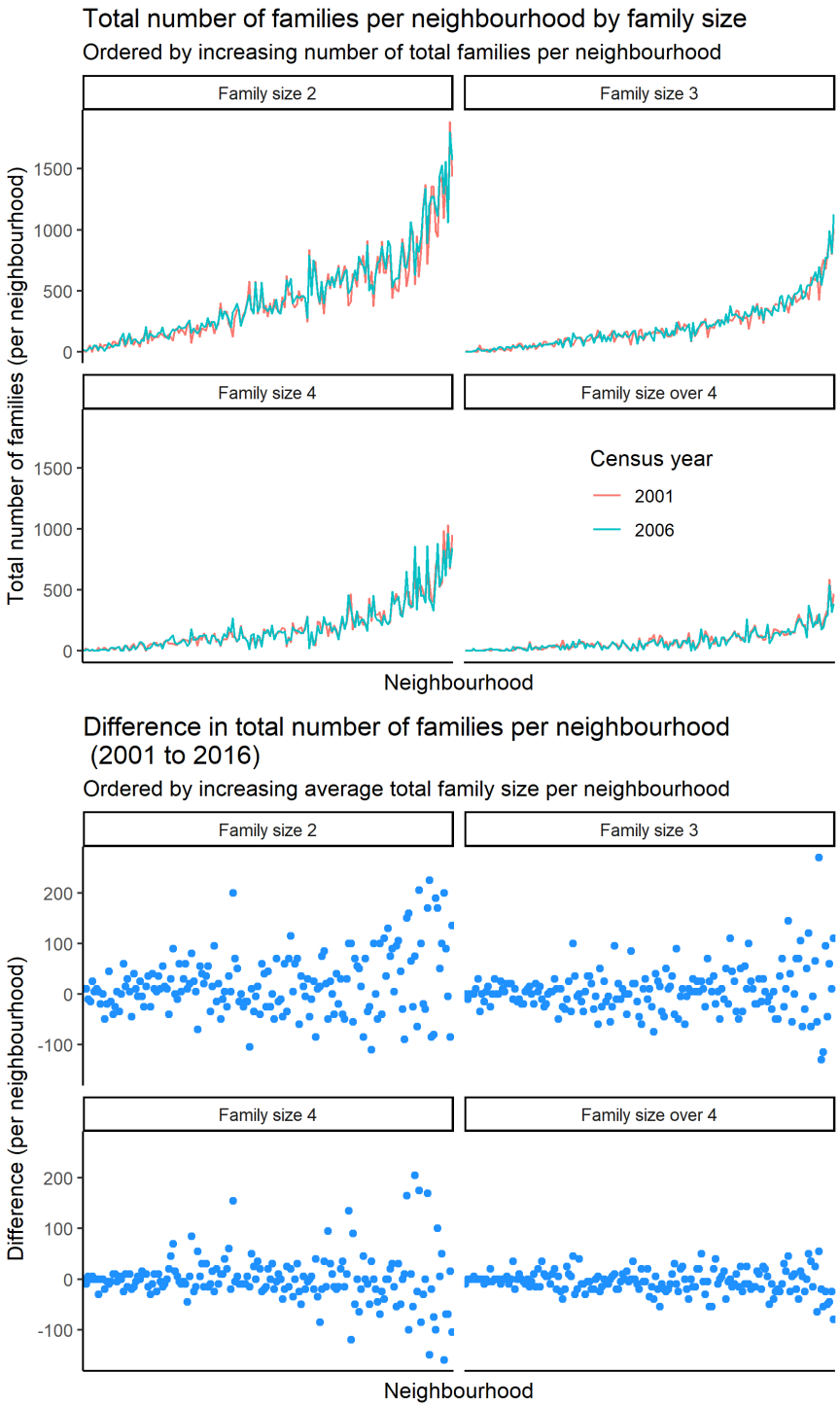

Figure 2: Top: Total number of families by neighbourhood separated into the four categories. Bottom: The differences are calculated as the 2006 value minus the 2001 value.

From Figure 2 it's clear that as the total number of families in a neighbourhood increase so does the total number of families for the four different categories. However, not all family sizes grow at the same rate. It appears that in general, as the total number of families grow in a neighbourhood the smaller size families grow more quickly then the larger families. Meaning that the difference between the number of families of size two and families over four will be much greater for neighbourhoods that have a large total number of families versus a neighbourhood that has a low total number of families.

From the graph above it does not appear that there is a lot of change in the family sizes for most nighbourhoods between 2001 and 2006. So we plot the differences and order them based on the average total number of families in a neighbourhood for 2001 and 2006 to see if there really is little change in the family size counts. We ordered the counts based on the average family size of both years combined since the yearly counts generally follow a linear relationship with slope close to 1, as was shown in Figure 1. So there should be minimal change in the ordering for the differences. From this it appears that as the average total number of families increase there is greater variance in the difference between the two years for small family sizes, but large families (those over 4 people) don't seem to vary much.

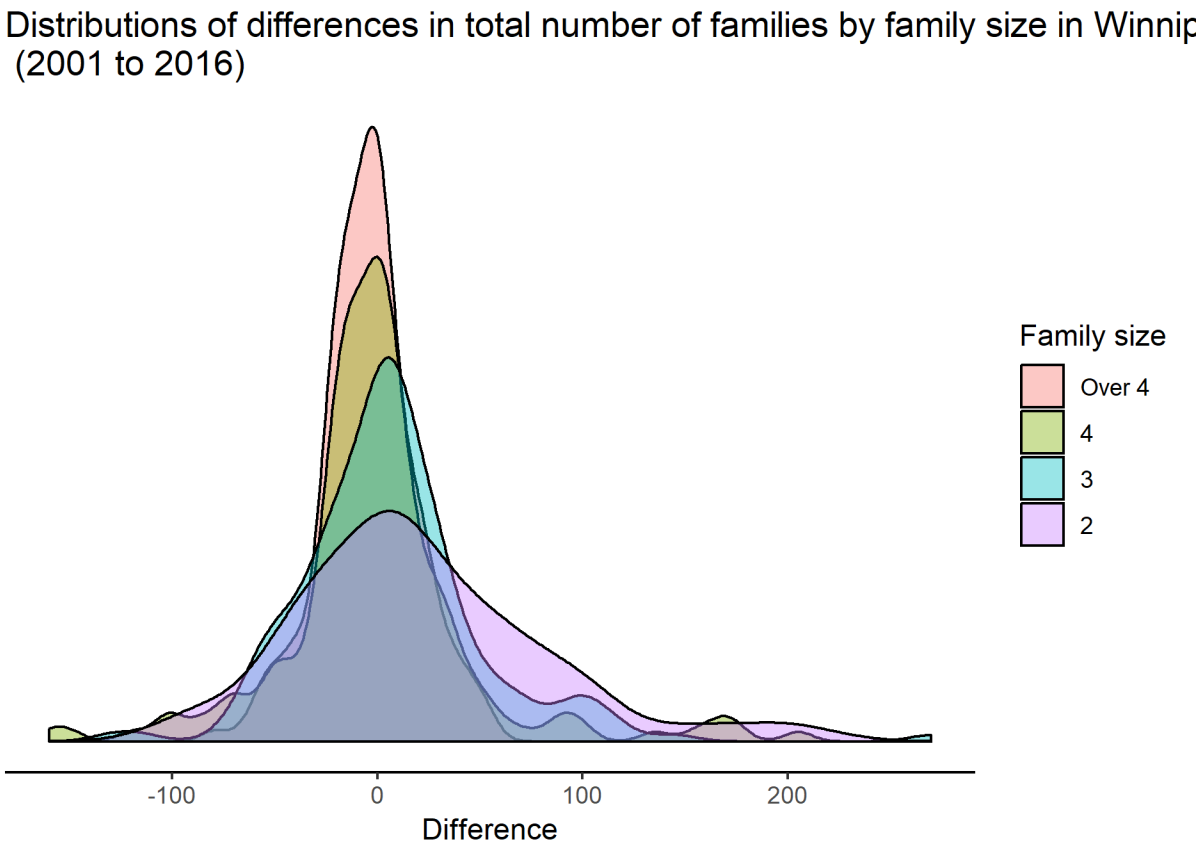

Figure 3: Combined density plots for the difference between 2001 and 2006 separated by family size.

In Figure 3 we show a combination of density plots showing the distributions of the differences for the four family size categories. It seems that the visual intuition from the previous figure is pretty close. Family sizes over four do seem to have the smallest variance, and families of size two the smallest, but families of size three and four need a closer inspection. Bartlett's test for equality of variances returns a p-value close to zero, further supporting the difference of variances. An assumption of Bartlett's test is normality for the populations being compared, and Table 1 contains the results of the Shapiro-Wilks test for normality, indicating this assumption is satisfied. The individual variances are shown in Table 2 and support what we've

concluded for the most part, except that families of size four have slightly higher variance than those of size three. So there may not be a continual decrease in variance as family sizes grow, but there is certainly a trend towards that.

Table 1: Results for the Shapiro-Wilks normality test.

Family Size	P-value
Size 2 Diff	0.000004
Size 3 Diff	0.000000
Size 4 Diff	0.000000
Size Over 4 Diff	0.002209

Table 2: Variances and Standard Deviations for the different family size differences between 2001 and 2006.

Family Size	Variance	Standard Deviation
Size 2 Diff	3636.8288	60.30613
Size 3 Diff	1988.7257	44.59513
Size 4 Diff	2252.9101	47.46483
Size Over 4 Diff	484.0693	22.00157

Conclusion

So the conclusions we can draw thus far is that though there is no significant changes for the total number of family sizes in the different neighbourhoods from 2001 to 2006, there are differences between the various family sizes in the rate of increase as the neighbourhood population increases. Also, the number of small families in a neighbourhood change more frequently than the number of large families. One possible explanation for this is that large families probably don't move very often. Which seems reasonable as families with a lot of people generally have a lot of children or take care of the elderly in the family. These two situations would appear to entail a lifestyle of consistency, meaning that larger families would not want to move very often.

Reflection

I learned that finding an interesting data set can be the most time consuming part of a project. This is because once you find a data set you need to do some initial exploration to verify whether it's an appropriate one or not. This step can't be skipped, but hopefully the amount of time spent here can be minimized. When the right data set is found, the analysis takes little time compared to the time spent looking finding it. I also found the analysis to be fairly enjoyable once I found the right data set, because there were some interesting things to say about it. I would recommend future students to start looking at data sets as soon as possible because procrastinating on this part can be hazardous to your mental health, and GPA, if it's put off to long. In terms of what I would have done differently, I would have had more sessions with the instructor to discuss ideas or talk about where I could find good data sets at an earlier point in the project.