36-315 Statistical Graphics and Visualization

Luke Davis, Michael Ferraco, Inez Foong, Jiunn Haur Lim

1 Population Distribution by Race

We examined the population distribution of race in 3 ways - population size, population proportion and, latitude and longitude distribution by population proportion. In 2000, we had tracts that had missing values or had a population of size zero. We decided to color the tracts white.

1.1 Map of Population Size by Race

In both the maps for 2000 and 2010, we see that whites tend to be clustered towards to west while blacks tend to be clustered towards the east. In 2000, there were 3 global modes for whites - northwest, central and eastern Louisiana while in 2010, the global modes in eastern Louisiana is much smaller. Hence we see that there has been a shift whites towards western Louisiana. In 2000, there were 2 global modes for blacks both in northeastern Louisiana while in 2010, these global modes were smaller but we see that the local modes in southeastern Louisiana showing an increase in the population of blacks there.

There was a tract in the south which had a large population of Native Americans (between 100-700) in 2000 but this decreased significantly in 2010 (between 10-100). Looking at the distribution of Hispanics, we see that it remained approximately the same. They continued to have a larger presence in the northeastern and central Louisiana. Finally, looking at the distribution of Asians, we see that in 2000, they used be in the northeastern parts of Louisiana while in 2010, they are now also in the northwestern and southern parts of Louisiana.

We decided to examine the population distribution of each race on the map as it would easier to visualize if those from a certain race tend to cluster together. However simply plotting population distribution on the map might make it harder to compare difference between races since there are races with extremely large populations such as whites and races with extremely small populations such as Native Americans. As such, we decided to include an option for the user to add population contour plots on the map to help the user identify clusters of tracts where there are large populations of each race. An alternative to using contour plots would be using heat maps. However, overlaying heat maps on the maps might make them difficult to read.

1.2 Map of Population Proportion by Race

For whites, see that in general, they have a large presence in most tracts or block groups as they make up more than 50% of the population of most tracts or block groups in both 2000 and 2010. We however see that like in the map of population size by race, there are more tracts in 2010 in the east with a smaller proportion of whites (< 50%).

For blacks, however, we see that they continue to be clustered in the eastern region of Louisiana. We also see that there are tracts in 2010 that now have an increased proportion of blacks in the southwestern tracts. For Asians, we see that there used to be a tract with a large proportion of Asians in the southwestern region of Louisiana in 2000. In 2010 however, we see that the proportion of Asians in this region has decreased. This seems to be the opposite of what happened to the blacks in this particular region. Hence it seems like the blacks started moving towards the southwestern region of Louisiana while Asians started moving out of this region. Finally, for Hispanics, we see that there is an increase in the population proportion of Hispanics in the northern tracts between 2000 to 2010.

We decided to examine the population proportion distribution of each race on the map as it would easier to visualize if those from a certain race tend to cluster together. Coloring the map by using population proportion of each race, helps to control for the differences in the absolute size of the difference population sizes. However we will still need to plot two side-by-side maps in order to visualize the change in population proportion which increases data ink. An alternative would be to plot a single map showing the percentage change in population proportion which would allow us to plot the same information but on single map. However since we do not have unique FIPs codes for each plotted tract in 2000, we could not calculate the percentage change for each plotted tract or block group. Also, just like what we did for our map of population size by race, we decided to include an option for the user to add population contour plots on the map to help the user identify clusters of tracts where there are large proportions of each race. An alternative to using contour plots would be using heat maps. However, overlaying heat maps on the maps might make them difficult to read.

1.3 Violin Plot of Latitude and Longitude Distribution by Population Proportion for each Race

Besides looking at the distribution of population proportions on the map, we also used violin plots to examine the distribution of population proportions. Just like what we saw in the maps, we see that tracts with a greater proportion of whites are located towards the west while those with a greater proportion of blacks are located towards the east. For Native Americans, we see that in 2000, tracts which were made up of 50% or more of Native Americans were located in the east and were located in a region that was different from tracts which had 20%-35% and 35%-50% native Americans as their IQR do not overlap. In 2010 however, we see that this is no longer the case. It seems that tracts which were made up of 35%-50% of Native Americans were located in the east. Hence it seems that some of the Native Americans had moved out from that region into other regions. Asians and Hispanics did not seem to have changes in the east west distributions between 2000 to 2010 as the violin plots look about the same across both years.

For our plot, we decided to use the default bandwidth to show the user what the distributions would look like given the different distributions. This is because we are looking at the distributions for both latitude and longitude for the different races. Hence setting a bandwidth that works particularly well for a latitude or longitude or for a particular race may not work as

well when it changes. As such, we added controls to allow the user to change the bandwidths to visualize how the distribution changes depending on the bandwidth.

The advantage of using a violin plot is that it allows it to examine the east-west and north-south distribution of population proportions separately but its disadvantage is that there is a need to choose bandwidths. An alternative would be to use boxplots. Boxplots would not require use to set bandwidths but would not allow us to visualize the distributions as well since it would only provide us with information about the 5 figure summary. Another alternative would be to use heat maps. However using the heat maps would require us to choose more bandwidths than what we would needed to for the violin plots. The advantage of using a heat map is that it would allow us to visualize the latitude and longitude distribution of each race on a single plot rather than over 2 separate plots hence saving data ink. However, we thought that it might be interesting to examine if there was a difference in the distribution of the races across either just the latitude or longitude. Additionally, the contour plots on our map of population distribution already indicate where the modes are with respect to population distribution across the races

2 Age and Race in 2010

To investigate the relationship between age and race, we used a 2D kernel density estimate of age and racial proportion - both are continuous variables - and visualized the results on a heat map and contour plot. This allows us to observe the 2-dimensional distribution and look for modes. As usual with density estimates, we have to check that the sample sizes are sufficiently large. Since there are 3471 block groups, this is not a problem.

Finally, to choose an appropriate bandwidth for the density estimate, we plotted the actual points over the heatmap and contour plot and manually adjusted the bandwidth, decreasing it until the curves became too sensitive to noise (too many local modes), and increasing it until modes start to disappear. Furthermore, we added a slider (using Shiny) to allow the reader to adjust the bandwidth where necessary.

From the plot, it is immediately apparent that there are two modes in the distribution with a large vertical gap, which tells us that there are some block groups with a large proportion of whites (mode A), and some block groups with a very small proportion of whites (mode B). Further investigation revealed that the latter block groups have a large proportion of blacks. This reveals a racially segregated community.

Moreover, mode A lies to the right of the vertical dotted line (which shows the average age over all block groups), whereas mode B lies to the left. This indicates a potential correlation, where block groups with higher proportion of whites tend to have a higher median age. Note that the average age (for both males and females) lie between 35 and 40, which is close to the average seen in other US cities.

Alternatively, I could have used the racial proportions to identify block groups that are predominantly white or predominantly black, and then created overlapping histograms or choropleths of the median age of block groups in each category. However, that seems to throw away too much information about racial proportions. In particular, there seem to be several block groups with between 40% and 60% whites, and so it may be unfair to simply toss them into two discrete categories.

3 Income and Race in 2010

As before, we used a 2D kernel density estimate of income and racial proportion to estimate the joint distribution of median income and racial proportions. This allows us to observe the 2-dimensional distribution and look for modes. Again, since we have 3471 block groups, sample size is not a problem. The bandwidth was chosen for the same reasons as before.

From the plot, it is again immediately apparent that there are two modes in the distribution with a large vertical gap, which tells us that there are some block groups with a large proportion of whites (mode A), and some block groups with a very small proportion of whites (mode B).

Moreover, mode A lies to the right of the vertical dotted line (which shows the average income over all block groups), whereas mode B lies to the left. This indicates a potential correlation, where block groups with higher proportion of whites tend to have a higher median income.

Switching the chart to show median income for females while fixing the horizontal scale, we see that there is a general shift to the left, indicating that females generally earn less than males. This phenomenon is seen across all US cities. However, the absolute difference between the two modes is smaller for females than that of males.

Alternatively, I could have split the block groups into two categories based on the dominant race, but that throws away too much information about the racial distribution. I could have also used a log scale for the horizontal scale (as is the convention when dealing with income), but in this case the range seems small enough to do without a log scale.

4 Age and Population 2010

To compare age and population, we decided to plot maps of New Orleans color coded by age distribution. The darker color of the block group indicates a higher age and vice versa. These maps have a checkbox above them where the user can select whether or not to show the overlaying population density contour plot. These maps were also broken up by gender, with the map on the left showing the distribution of age for males and the map on the right showing the distribution in age for females. The population density contour on each map shows the population density for both genders.

The large white area to the north is water as well as the white area to the northeast. The white area to the northwest is the airport and the white rectangle in the north in the center of the map just above the red square is a park.

When first looking at these maps without the population density contour plot overlay, there are a few findings worth noting. The male and female graphs both show that the outskirts of the city seem to be made up of older people than more towards the center of the city. Although the males seem to be a little younger in the southeast of the map, in general, both genders seem to be older outside the city center. Also, on both the male and female maps, there is a narrow strip of younger people in the block group in the center and to the south.

The most important observation is that directly in the center of the city, there is a majority of younger people in both genders.

There are a few differences in the age distributions for each gender. One notable difference is that the males seem to be younger in the northeast than the females.

When overlaying the contour plot, we can see that there are three very distinguishable modes. These modes are present in the direct center of the city along with the northwest and northeast. There are also some smaller, less distinguishable modes, just south and to the east of the center of the city.

The densest population, when looking at this contour plot, is in the direct center of the city. This area is made up of young people as mentioned earlier and as seen when looking at the age distribution by color. Below the maps of New Orleans, we have included zoomed in versions of this densely populated area to better see the age distribution.

We chose to display this information on maps zoomed in on New Orleans. We thought that these maps offered the best visual representation of the data. Using the map to plot this data allowed us to use the location statistics to better help our reader understand the data. If we broke down age in each block group by using some other form of data such as a scatter plot, we would not have been able to compare it with population in such an easily comparable way.

A disadvantage of using this type of representation is that it can be difficult to view some of the block groups in the city. When moving into the densely populated areas, the block groups become much smaller and it can be hard to look at these areas and compare them. This is the reason the zoomed in versions of the center of the city were included.

In addition, we offered the zoomed in versions of the direct center of the city to allow for an easier look at where our most important statistical finding was. It makes it easier for the reader to look at this part of the city and understand our reasoning behind our conclusions.

5 Income and Population 2010

To compare income and population, we decided to plot maps of New Orleans color coded by income distribution. The darker color of the block group indicates a higher income and vice versa. These maps have a checkbox above them where the user can select whether or not to show the overlaying population density contour plot. These maps were also broken up by gender, with the map on the left showing the distribution of income for males and the map on

the right showing the distribution of income for females. The population density contour on each map shows the population density for both genders.

The large white area to the north is water as well as the white area to the northeast. The white area to the northwest is the airport and the white rectangle in the north in the center of the map just above the red square is a park.

When looking at these maps without overlaying population density contour plots, we can see there are a few findings worth mentioning. Outside the center of the city, on the outskirts of the map, there seem to be higher income people than closer to the center of the city.

Also, if we look at the zoomed in versions of the direct center of the city below, we can see that there seems to be a low income population in this area.

The population contour overlaying plot is the same as the one for the age distribution mentioned above. An important area to look at again is the center of the city, which is zoomed in below these maps, due to the high density population and interestingly low income distribution.

The densest population, same as above for the age distribution maps, is in the direct center of the city. This area is made up of lower income people as mentioned earlier and as seen when looking at the income distribution by color. Below the maps of New Orleans, we have included zoomed in versions of this densely populated area to better see the income distribution.

We chose to display this information on maps zoomed in on New Orleans. We thought that these maps offered the best visual representation of the data. Using the map to plot this data allowed us to use the location statistics to better help our reader understand the data. If we broke down age in each block group by using some other form of data such as a scatter plot, we would not have been able to compare it with population in such an easily comparable way. Comparing this data was very similar to comparing age and population, so we decided to make similar graphs for both data comparisons.

A disadvantage of using this type of representation, similar to the age vs. population maps explained above, is that it can be difficult to view some of the block groups in the city. When moving into the densely populated areas, the block groups become much smaller and it can be hard to look at these areas and compare them. This is the reason the zoomed in versions of the center of the city were included.

In addition, we offered the zoomed in versions of the direct center of the city to allow for an easier look at where our most important statistical finding was. It makes it easier for the reader to look at this part of the city and understand our reasoning behind our conclusions.

After comparing age with population and income with population, a few important findings stand out. As mentioned above, the younger people seem to live in the center of the city while the older people seem to be outside this city center. Also, when looking at the income maps, as mentioned earlier, there seems to be lower income people living in the center of the city, while the higher income people are living outside the city center. Again, as mentioned earlier, the population density is highest in the center of the city and is very low on the outskirts.

Due to these conclusions, we believe that people with established careers and families that are higher in age live outside the city center on the outskirts of New Orleans. However, younger people seem to like to live directly in the center of the city and these people have lower incomes due to their age and lack of experience.

6 Barplot of Racial Breakdown of Louisiana

We chose to use two barplots on top of each other to show the difference in the proportion of each race between 2000 and 2010. The colors, order, and axis scales match on the graphs to make it easy to compare. We manipulated the margins on the two plots to get them right next to each other to reduce the distance the viewer has to travel with his or her eyes. We chose colors for the bar to be as distinct as possible.

After seeing that whites and blacks together made up over eighty percent of the population, which made changes in the proportions of minorities between the two census years difficult, we decided to add an option in shiny to exclude whites and blacks from the graph. In this second graph the increase in Hispanic and Pacific Islanders is much more obvious due to the greatly reduced x-axis scale. We kept the colors assigned to each race consistent across the two graphs to minimize confusion.

7 Barplots of Household Size by Race

We experimented with using a dotchart to fit all the data on a single plot but decided that the display was too crowded and wanted to use a more familiar display. We used a single barplot for each race since having all four races overlaid was also too busy. Switching between two races in Shiny makes it pretty clear which household sizes make up a greater or lesser proportion of all households for that race. We chose to use proportion of households within each race to make comparisons possible. Since the disparity between the representation of different races in the state is so large doing any kind of frequency count would have made comparisons between graphs much more difficult. Purple was chosen as a compromise between red and blue.