Joe Pane, Julian Zhou, Chris Conte, Matt Goldberg
36-315 Group 9 Final Project Report
Due: Thursday, May 9

1) These charts are meant to show that the Memphis, West Memphis area is made up primarily of blacks and whites and that the proportion of blacks is growing, while the proportion of whites is shrinking. To show the distribution of race in Memphis and West Memphis, spine charts, bar graphs, pie charts, and rose diagrams can be used. Because proportion is more interpretable than counts, the axis of a bar graph should be changed. The bar graph is easy to interpret exact counts, though the message of the plot does not require specificity. The spine chart can be difficult to compare, though would be feasible for the key point. The rose diagram does not provide measures of exact counts or proportions, but is faster to compare than the pie chart. The advantages of the rose diagram are that it gets across the key message and is pretty. However, there are some disadvantages of a rose diagram and that is the difference in proportions between blacks and whites is understated in the rose plot, because of the geometry associated with circles (as the radius increases, the area of the circle/pie slice increases at an even faster rate.

   Comparing the rose diagrams over time, we see an increase in the amount of blacks and minorities who identified themselves with 2 or more other races, as represented in black and orange respectively. Before introducing our choropleths, it is worth mentioning that West Memphis, AR, occupies around 2/5 of the plotted area on the left, while Memphis, TN, occupies the other 3/5 area on the right. In the choropleths we observe the same trend that the black population is increasing while the white population is shrinking. This effect is strongest in West Memphis, as indicated by the fact that a few block groups in West Memphis who were colored

green in 2000 (indicating higher white population) became brownish in 2010 (indicating lower white population), while many block groups in West Memphis that were colored brownish in 2000 became yellowish or green in 2010. Another way to show this change would have been to plot the difference in proportion or counts for each race over time. However, this would not have provided information regarding the relative proportion of each race in Memphis and West Memphis. Furthermore, without a map it is very hard, if not impossible, to represent the spatial component of the changes in population that took place over time.

2) The main point of the boxplot is that females tend to be slightly older than males, and that the average of both male and female in the Memphis and West Memphis area is about 35 to 40 years old. Disadvantages of the boxplot are that it does not show information regarding modality or shape, whereas a violin plot, bean-plot, and box-percentile plot could have all shown the distribution of age, and provided more detail as to the modality and shape (though it is generally assumed age is normally distributed). Also, since this data came from the Census, we know that the population is represented by all of the people in Memphis and West Memphis. Therefore, we are not just getting a random sample from the population but we are actually getting the whole population. If age was found to follow an abnormal distribution, the information regarding its shape would have been valuable, but that requires using one of the alternatives abovementioned rather than the boxplot. The advantage of the boxplot is that it easily identifies the outlier of interest: block groups containing average male age and average female age that are over 80. In particular, if time permitting, it would be interesting to see if the two outliers over 80 for both males and females came from the same block group. In addition,

suggestions to the Census for future work would be to see where nursing homes, universities, etc. are located. With this potential additional information, we could make even more conclusions.

The second plot is extremely noisy and somewhat overwhelming at first glance. Examining it from a further perspective it is very clear that females are generally older than men (the pink is higher than the blue blob). The use of X's and O's is valuable to those who can not see color, while not disadvantaging those who can as the same amount of ink is used in each shape. A heat map and contour plot could have shown the same information, but they take slightly longer to interpret, and given the key message—this level of detail was not necessary. In addition, there would have been much more ink on the heat map than was actually needed to interpret the simple story from this graph.

3) In examining income by block group in Memphis and West Memphis, it was determined that males tend to earn more than females and had a far wider range of salary. The discrepancy is far larger than the separation of any other variables into most categories, and so a graphic that would best show this large discrepancy was used. The density plot is powerful in that it is immediately apparent that males earn far more than females on average. One possible explanation to this could be that men have many more career options available to them, assuming that no businesses are illegally paying females less for the same jobs. Boxplots, box-percentile plots, and violin plots, could have showed this information, but the density curves on the same plot best reflect the second point regarding career options. However, a box plot or violin plot, which clearly mark the center of each distribution, would have been more useful to

show the first point. In balancing the relative value of each point and the plots abilities to show each point a density plot was selected.

To examine any possible correlation between income and block group total population, a contour map on top of a heat map were used, each representing female and male income respectively. Unsurprisingly, the block group total population levels at which densities of female and male income are the highest appear to be the same. This is because whereas block group average income is categorized by gender, block group total population is not. This intrinsic limitation of the raw data to a certain extent limited the amount of useful and interesting information that can be explored and conveyed by the graphs. For instance, if possible, it would be more interesting to compare how block group female income correlates with block group female total population, with how block group male income correlates with block group male total population. Furthermore, at a given block group total population level, the income level at which 2D density is the highest is highr for male than for female. This is indicated by the fact that the spot with the greatest red intensity in the heat map is vertically higher than the innermost contour line. Last but not least, it is noteworthy that while the contour map suggests that female average income tends to occupy a range from 0 to 50,000, male average income tend to vary much more widely, from 0 to as high as 110,000, as indicated by the lighter blue region in the heat map. This difference in male and female income range is generally consistent with the trend indicated by the 1D density plot. This graphic takes longer to interpret, but provides more insight into the quantitative values of the effects portrayed in the first plot. This could have also been done reversing the contour and heat map for each gender without the loss of much information. The choice was fairly arbitrary, except that the range was larger for

males and the use of colors in the heat map made it prettier. In choosing the amount of contours, the balance between the amount of noise on the graphic and the clarity of the point was considered. The amount of contours and sensitivity of the heat map was chosen so that the point would be clear, with the least amount of noise. Points could have also been used to compare the heat map and the contour map, though this would add a lot of noise and make the genders harder to compare.

4) These four perspective plots (with scaled axis for easy comparison) demonstrate that block groups tend to have smaller black population in general compared to white population, and that the average age in these block groups tend to be more variable (in addition to re-emphasizes the point made in the second set of graphics that females tend to live longer than males). Because of the consistent scale of the axes, it is immediately obvious that the ages of black males are more variable, because the mode is not as high, and the lack of spread ("pointiness") of the mode for whites is distinguishable. We also see that the mode for females (both black and white) is generally pushed to the right indicating that they live longer. The colors used do not indicate anything, though they could have been useful to indicate gender or race. However, this would limit the takeaways of the graph to be along gender or racial points, while much more information is displayed.

   These four plots are useful in that they quickly summarize many points like main racial and gender combinations on the distribution of age. However, the lack of numerical labeling on the axis and complexity of three-dimensional graphs does not allow for detailed, quantitative takeaways. This could have been done with two dimensional graphs examining key

relationships as used in series 2 and 3—however this would limit the amount of takeaways, and perhaps leave unanswered questions for the viewer. With that being said, our previous graphs gave the viewer information about the quantitative data portrayed in these graphs. Because of more specific charts used in other series, this series used the four graphs to give an overview of age over two variables.

5) After the takeaways of the first four graphs, the viewer is undoubtedly wondering about the difference in income along racial variables. As shown in the pervious graphics this could be done with boxplots, box-percentile plots, density plots, and violin plots across each category or a heat map and contour plot—or with the use of points in a scatterplot. By using points in a scatterplot, the effect is not as clear, though a general difference can be noticed. In the previous income comparison a heat map with an added contour plot had already been used to hit home the large discrepancy between genders. There is still a very large discrepancy among the races, and so two heat maps with the same scales on the axes were used to show the large differences in income between races. The disadvantage of this is that the takeaway is now spread across two graphs, whereas in the previous series, the viewer could understand the main point in the first graph and was provided with more detail in the second. The advantage is that no techniques were repeated and the magnitude of the differences in income is comparable.

   From the side-by-side 2D KDEs, the axes of which have been standardized, we see that the income-population correlation for whites and blacks are quite different. For whites, density tends to be highest where male income is around 18,000 to 30,000 and where block group

white population is around 0-400. For black, density tends to be highest where male income is similarly around 18,000 to 30,000 but where block group black population is around 500-1,200. In general, black male income-population tends to have more variability than that of the white, as indicated by the heat map for the blacks having a larger area of the heat map where varying colors are observed. Female income-population correlations for whites and blacks largely agree with male income-population correlations, as indicated by the fact that density of the female scatterplot points tend to agree with density of the heat map.

6) In describing and showing how the distribution of households of different sizes looks like, we first examine the violin plot. This shows us household size and how many of each type we have. Immediately noticeable is the fact that very few houses have 6-7 people, it is significantly less than the other categories. There are also more 1-2 person houses than 3-5, the 75^{th} percentile of 1-2 person households just about matches up with the end of the main body of the violin plot of 3-5, while the outliers both stretch out across the x-axis.

   We chose the violin plots for this data because it shows the most information for this situation. This is because it gives us the same information that can be shown by a boxplot, plus the density curve around it. For this reason, it is better than just a boxplot or density curve by itself. We could have also used a series of bar plots but this would have also lost information for us because that would have only given us counts and amounts, not the specific quartile values or density information that we are looking for.

   Finally we have choropleth maps, which show how the distribution of house size varies across regions. Immediately, we see that the average household size in W. Memphis seems to be

smaller than in the Memphis area. This could be because there are more people in Memphis and more young people living on their own. The maps also confirm the finding of the violin plot, that there are many more 1-2 person houses than the others.

This map complements the violin plot because it adds more substance, mainly geographical information, to the data. Now we have a better idea of where the different sized houses tend to lie throughout the metropolitan area.

7) In order to show if there is a correlation between household size and age, we will take a look at some 3D perspective plots. These plots show the average age of persons by gender in a block group in relation to the proportion of households of a certain size in that block group. For both genders, 1-2 person households tend to be older than 3-5 person, and then 6-7 has much younger. This makes sense. One would imagine nobody is really living on their own until at least age 18, and in most cases much older than that. A house with lots of people living in it could be for a variety of reasons, but in many cases it is because of numerous children, and they make the average age much younger. As well, lots of seniors live either by themselves or with their husband/wife, so it makes sense that 1-2 person households would be oldest category. Women seem to be older in 1-2 and 3-5, but it is hard to see much of a difference in the highest category. This could be because children are the same age regardless of gender, and because there may be a smaller sample of 6-7 person homes. Women may be older overall in those categories because they have a higher life expectancy, so their average is slightly shifted.

   Advantages of the perspective plot are that we can easily see particular stories on the graph that we predicted based on previous plots. They are also pretty to look at and can be rotated

to see different parts of the graphs. Disadvantages are that we have no sense for numerical measure on the graphs and we also cannot adjust theta, the degree of rotation, on the graph without rerunning the code and re-plotting the graphs. Shiny could have been something to incorporate in our graphs to make it easier to adjust theta right in front of the viewer. Other options we could have used include heat maps and contour maps – the usual 2D kernel density estimate visualization. In fact, these two might have been even more advantageous in that both provide numerical measures along the axes. The decision to refrain from using heat maps and contour was simply to avoid excessive repetition.

8) This is a map of Memphis/ West Memphis color coded according to average income and broken up by males and females. By directly comparing the genders with no change in range for categories, it allows us to easily see not only which areas tend to be more prosperous, but also how individual blocks change between male and female. One thing that is clear is that females have much fewer blocks in the darkest category, which correspond to higher income. That makes sense because we know from the previous graph that they make far less as a whole. Looking at the male graph, it seems that the richest area, represented by darker color, is in the north to north-west of the city center. The inner city seems poorest. West Memphis seems to be a bit better off but it is hard to tell. Certainly the western area seems more prosperous.

   It makes sense to display this data as a choropleth map because it allows us to see how different regions vary in terms of wealth and lets us see how this is different for men and women. As per the suggestion of Professor Thomas, we changed the colors of this and graph #9 to be uniform for the ages, as it makes direct comparison easier. Previously, it was hard to tell

which is the third darkest blue as compared to the third darkest red. By using the same color for both male and female, this problem has been solved. Advantages and disadvantages of the choropleth maps were discussed in series 1 and 4.

9) This is a plot of Memphis/ West Memphis color-coded by age. Examining the plot, we see that certain block groups make up what can be seen as an older or younger area. In particular, the southwest corner of Memphis seems to have somewhat older females living there, while the center and northeast corner tend to look younger. The outskirts and West Memphis area tends to change in no discernable pattern. Males seem to look younger overall, or at least have fewer blocks with the darkest shade. It would be interesting to see what proportion of block groups has the same color across gender category.

   We chose choropleth maps to show this data because it gives a clear picture of how age is distributed by geographic location across the city. A heat map and a contour plot would not have worked here because they are both 2D density estimates, which require one more variable. In this case we are only looking at one variable, age.

10) In this plot, we see first that the poverty level in Memphis/ West Memphis is overall higher (around 20%) than the national average (15%). Another key takeaway is the fact proportion of those living in poverty tends to decrease with age across all racial categories. This makes some sense; people who are younger tend to have more unstable economic circumstances. By race, we also see that blacks tend to be more impoverished than whites at all levels.

Here, we chose bar plots to represent poverty levels because we felt that it was the most straightforward way to tell what is perhaps the most important result of our findings. Memphis is a city in poor economic condition, and it's easy to see this by looking at the bars. Memphis is over 50% black, so the fact that the blacks of Memphis are so impoverished does not look well upon the city. A mosaic plot would have been a bit much here, and it would not have captured the subtle but substantial differences in proportion. Similarly, pie charts would not work well here because we are directly comparing percentages that do not add up to 100%.