This document is my (Daniel Paget) submission. The avenue I applied through was emailing and phone conversations with Campbell Pratt of Kleboe Jardine Performance hiring in science and technology.

Part (a) Texas with 1935 5-digit zip codes

Part (b)

(i) zcta5 = 04631 (Maine)
(ii) zcta5 = 96769 (Hawaii)
(iii) zcta5 = 56711 (Minnesota)

Part (c) zcta5 = 10028 (New York) with each house having 28.9 m^2 of land area

Part (d)

(i) 3-digit zip codes 834 and 063 are common to Wyoming + Idaho and New York + Connecticut respectively. Also, Hawaii contains zips with 3-digit zip code 968, but some entries in the dataset have ‘nan’ as their state value while also having 3-digit zip code 968. This is not a legitimately different entry as the 5-digit zip codes with this property have all the same values in their columns as entries within Hawaii. This means they are some kind of duplicate and not different entries.

(ii) There are two cities (Fisher's Island and Alta) for which the residents could easily be identified. These are the only two cities which have zip codes that do not match the rest of the state they are a part of (in the first three digits), and so if a resident of these cities are part of a record containing both their state and three digit zip code then they can be easily identified as living in their respective cities. Combined with sex age and race these individuals are very vulnerable to re-identification.

(iii) One of state or three digit zip code should be omitted from the dataset. It is more informative to include the three digit zip code, it gives more specific information about a patient's location, and on its own is not unsafe to provide. Therefore, the three digit zip code should remain in the dataset and the state should be removed to maximise the dataset's value while minimising risk of re-identification.

Part (e)

(i) 13 3-digit zip codes have populations of less than 20,000 residents.

(ii) Pros:

-Allows the data to be used in some capacity rather than discarded.
-Protects the participants from re-identification by removing the specific location information.

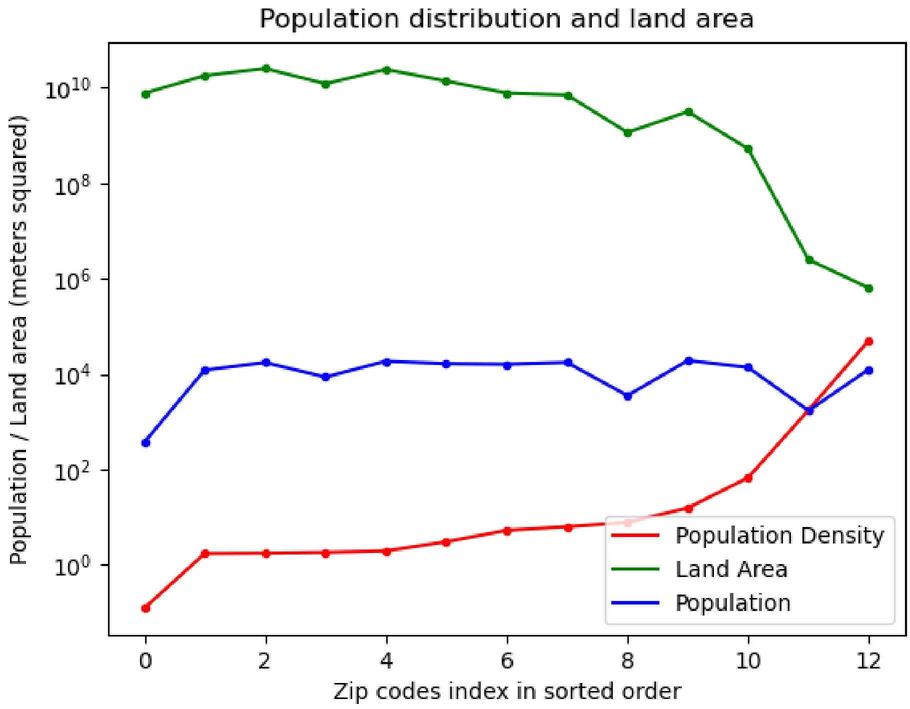
Cons:

-Lose the locational information of the patients put into group 000.
-If there are only a few 'small' zip codes then it is still potentially possible to re-identify the participants using additional external datasets.

Part (e)

(iii) Algorithmically merge all the patients belonging to 'small' zip codes into neighbouring zip codes. This should be done so that ideally their state doesn't change. This will preserve all of their data only smearing their location slightly. It also completely removes the risk of re-identification. This process should be done while moving entries to further and further neighbouring zips so that no entries remain with less than 20,000 residents.

(iv)



The above plot shows that ‘small’ zip codes, when ordered from lowest to highest population density (defined using housing and land area), the zip codes show no net meaningful change in actual population, but instead show a very significant change in land area. The smallest zip (by area) is 10,000 times smaller than the largest zip, yet has almost the exact same number of residents (in magnitude especially).

(v) The zip with the smallest pop density is 821 in Wyoming. This makes sense as there is only one 5-digit zip code with this 3-digit zip value, and so it is capable of being very abnormally high or low in population density. In this case, zip 821 has a population of 369 but a large area, and so has very low population density.

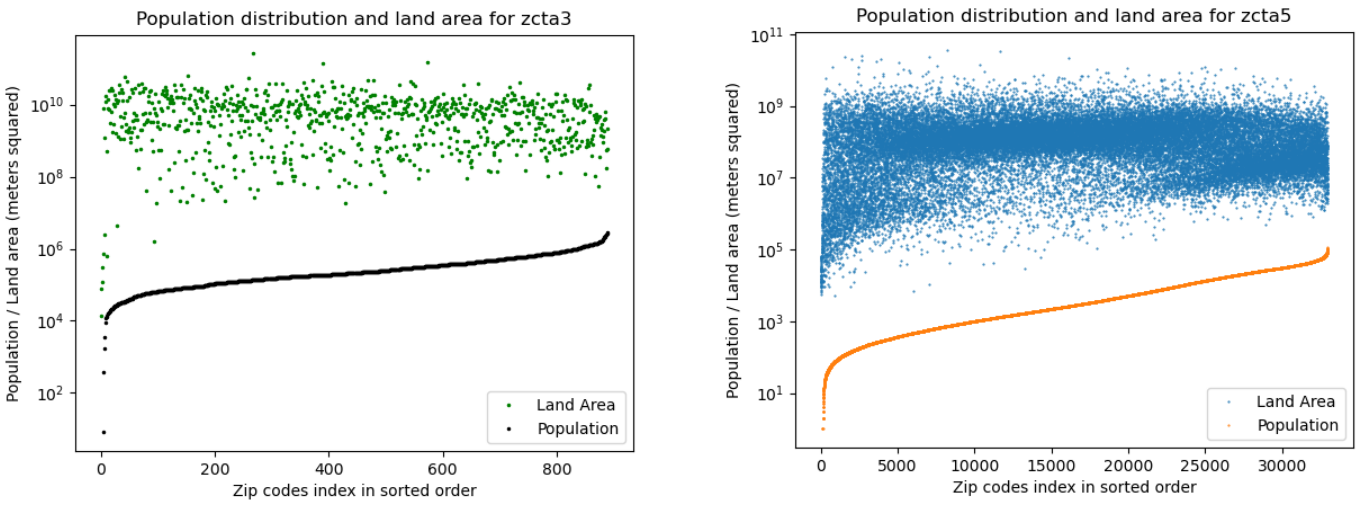


Figure A – (LEFT) The distribution of both the total population and total land area of each of the 3-digit zip codes. (RIGHT) Distribution of population and land area for each of the zcta5 (5-digit zip codes).

Plotted in figure A are the land areas and populations for all 5-digit zip codes and 3-digit zip codes. The above figure shows that the distribution of land area has multiple notable patterns, and the population is distributed in a very polarized way. Almost all 3-digit zips have populations in the region of 10^5 and 10^6 residents, however around 100 3-digit zip codes have less than 10^5 and a couple have significantly smaller populations. Both 3-digit and 5-digit zips show a small uptick at the end, which is consistent as one is a cumulation of the other so large spikes should align.

Most clear from figure A is the fact that the distribution of land area for 5-digit populations shows a significant slump from around index 0-6000, indicating that the lower populated 5-digit zip codes belong to the smaller areas of the US. Additionally, on the higher end of the population scale, the 5-digit zip codes with 10^5 residents have more residents than many of the 3-digit zip codes despite the fact that the population of the 3-digit zip codes are sums of many 5-digit zip codes.

Finally, from the 5-digit plot there are two interesting bands of data, one at around $10^8 m^2$ and the other from $10^5\ m^2$ growing to $10^7 m^2$ for the latter indexes. However, there is a gap in this band between index 10,000 and index 20,000 where there are far fewer zips with these band values. This is potentially further evidence for the polarised population, as this gap would indicate polarisation of population *density* for the large and small population zips. Interestingly this seems to be purely a 5-digit zip code phenomena as the 3-digit zip codes don’t show this pattern.

Overall, it is clear that there is a logarithmic distribution of population across the zip codes, both for 3-digit and 5-digit zips. There is a correlation between land area and population, and so a correlation between population and population density.