***Is there a correlation between diabetes and population density?***

**Team 5:** Jose Martinez Carvajal, Peter Kim, Mattieu Dubois, Ashley Peterson

LINE UP FOR PRESENTATION:

**Matt:**

Introducing the Question: ***Is there a correlation between diabetes and population density?***

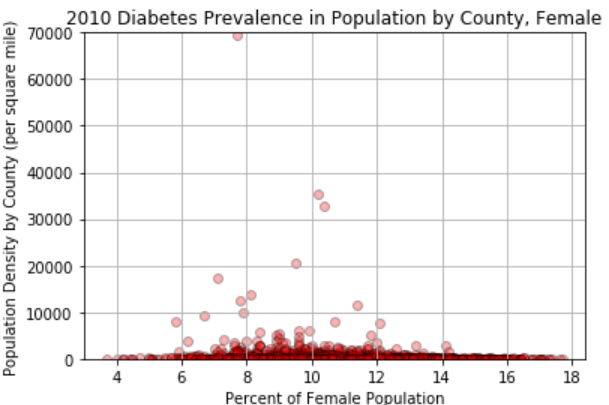
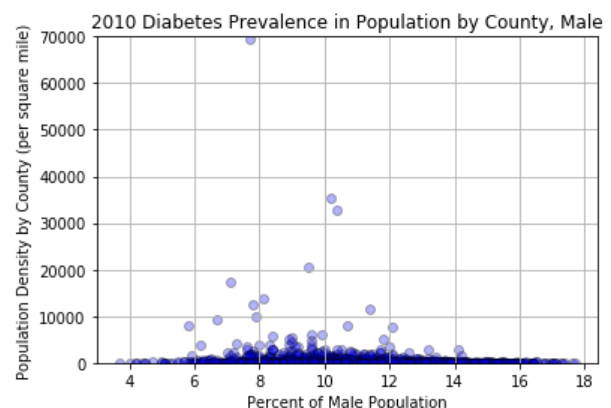
* Describe Data Clean up & Exploration Process
* Walk through **why** we chose the topic
* Were we successful?

**Jose:**

* Gender Analysis & T-Test

ASHLEY’s and Jose’s WRITE UP:

***Is diabetes more prevalent in one gender across the counties?***

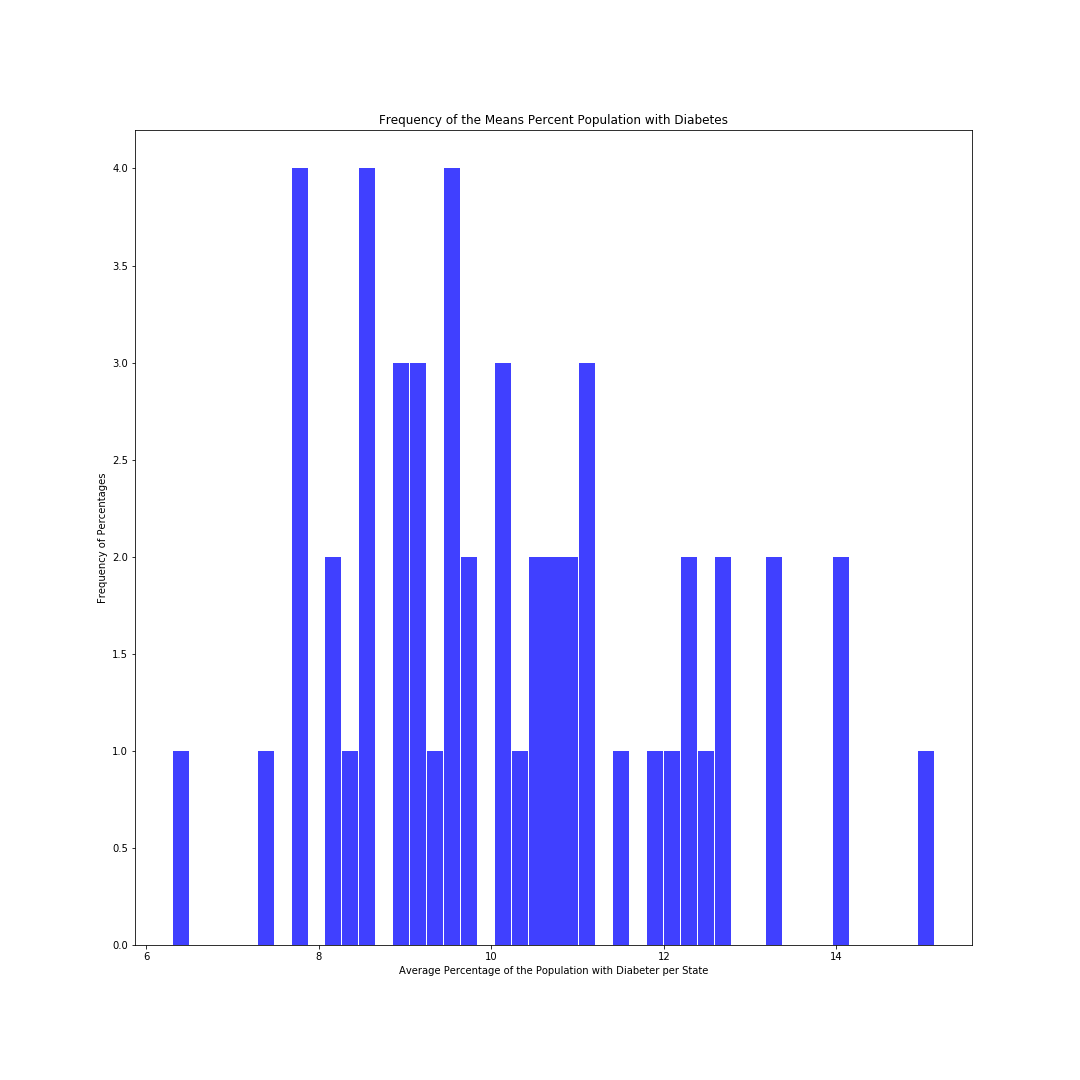


* We found that in comparing the 2010 census data with prevalence of diabetes in population by county, there was very little difference between gender shown in the scatter chart.
* The highest percentage of men with diabetes is located in Clay County, Indiana with 17.7%.
* The highest percentage of women with diabetes is located in Lowndes County, Georgia with 21%
* Average of diabetes across both male & female is largest from 8%-12%

|  |  |
| --- | --- |
| T- statistic | 16.2 |
| P-value = | 6.18e-58 |

* However, when we run the independent samples t-test, we see a high t-statistic and a very low p-value. This suggest two possible interpretations – the populations being studied are different from one another. However, the very low p-value points to another error. One of the parameters of the t-statistic is that both populations have no influence on each other. The p-value suggests that there is an influence between both populations. Since, the population consists of male and female, we may make a far fetch assumption that within the data for smaller counties there could have been husbands and wives, friends, and other relationships that could have an influence on one another.

Jose’s write up



* Using a list comprehension, we took the percentage of people with diabetes for all the counties per state and calculated the average. This gave us the average percentage of people with diabetes per state.
* Using these values, we constructed a histogram to look at the distribution of the average percentage population with diabetes.
* According to the distribution, the average percentage of people with diabetes across most states ranges from 8%-12%.

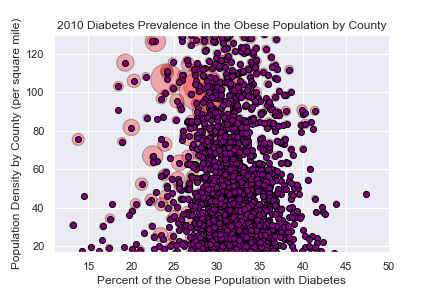
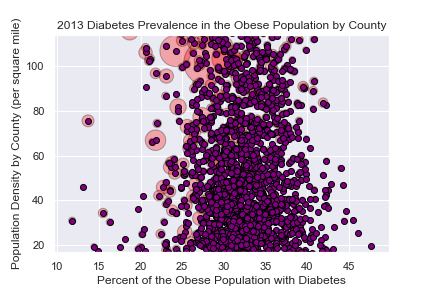
**Peter**: OBESITY and Diabetes

**Observations:**

1. Percentile pop:

The visual observation is supported by Pearson correlation value of -0.149 in 2010 and -0.151 in 2013. This a indicates a relatively weak negative correlation between the percentage of the diabetic obese population and population density.

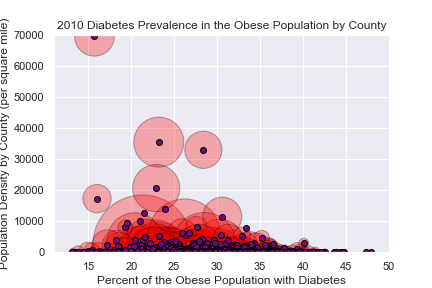
If plotted, a trendline would be leaning slightly backwards.

Between 25th and 75th of populations density by square mile.  
plotting indicates a slightly higher prevalence of diabetes in more sparely populated countries.

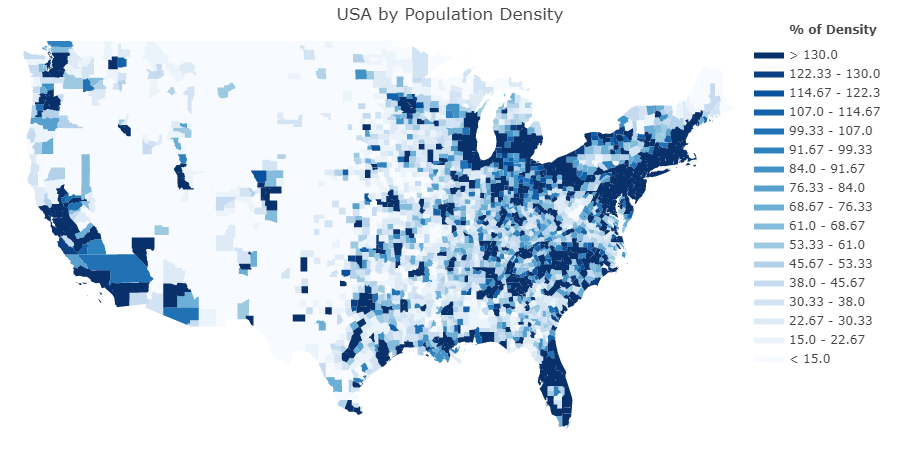
1. Full Population:

The pink halos represent the total population in the country. For example,

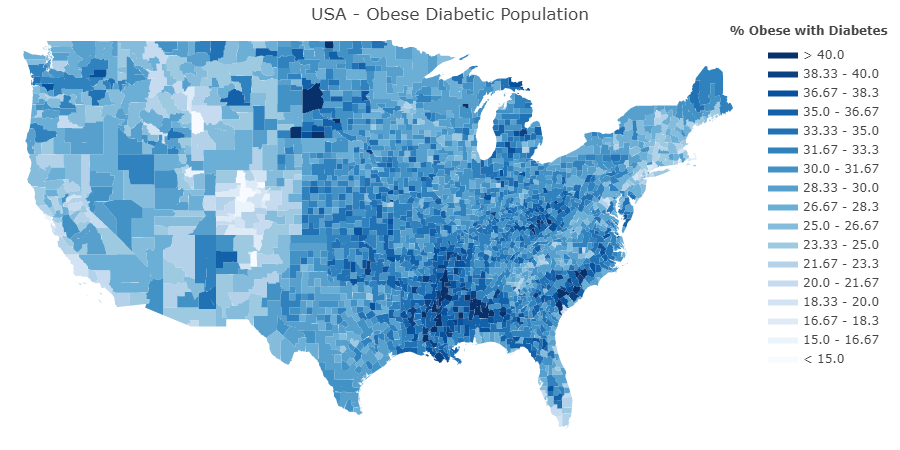
New York County (Manhattan) as expected is the most densely populated county in the US with of 70k heads per square mile. (total pop 1.5mil; 19 other counties have higher populations) (NYC Obesity diabetes prevalence of 15.7%)

* 1. Additional supporting link : <https://commons.wikimedia.org/wiki/File:New_York_Counties.svg>
  2. However, the largest county by population is Los Angeles 9.8mil but with a density of only 2.4k. (Obesity diabetes prevalence of 21%)

1. Heat Map – Population Density: For visual purposes, the colors were scaled to the 25th and 75th percentile of the data. Using the percentile in the prior scaled Bubble scatter plot representing the majority population, the color scale was also adjusted starting

from the 25th percentile to the 75th percentile for maximum contrast.

1. Heat Map - % Obese Diabetes: This chart shows the percentage of the obese diabetic population by county. Like the pop density heat map, the color scale was adjusted for contrast but this time visually between 15% to 45% marks.
   1. Observations: It was not a part of our hypothesis but I had assumptions of the areas in the country would be prevalent. In this case the heat map laid out that assumption.



**Hypothesis:**

The diabetic obese population in America reside in counties with a population density per square mile that is more than the median value.

The data for two periods was observed – 2010 & 2013. Both times periods would show a evidence of the null hypothesis.

Visually and through P value calculation, there was no correlation with the diabetic obese population residing in counties with a density below the national median.

|  |  |  |
| --- | --- | --- |
| Percentile of Population Density per square mile by County | | |
|
|  | **2010** | **2013** |
| **1st Quartile** | 17.6 | 16.9 |
| **2nd Quartile** | 46.6 | 45.2 |
| **3rd Quartile** | 129.4 | 113.8 |