Analysis of US Obesity using Visual Analytics

By Lesley Dwyer

Motivation, Data and Research Questions

Obesity is an epidemic in America and rates of adult obesity continue to increase [5]. It leads to chronic diseases like type 2 diabetes and heart disease and costs the US billions of dollars in medical costs every year [1].

Previous research has evaluated the effects of the neighbourhood and community on obesity in addition to individual factors. One US study found that living in a more disadvantaged community was related to having a higher Body-Mass Index (BMI); obesity is defined as having a BMI of 30 or more [8]. Another study in Canada found that neighbourhood-level factors played a part in obesity by encouraging physical inactivity and a poor diet [6]. Previous research also indicates factors associated with diet, exercise, education, income, race and gender are linked to obesity at an individual level [1, 3, 6, 8].

I plan to take a visual analytics approach to investigating obesity in the US by county. I will look at variation in obesity rates using the 2018 County Health Rankings compiled by The County Health Rankings & Roadmaps program [2]. This dataset includes several health, socio-economic, environment and demographic factors measured for each county from 2010-2017, including obesity. I will combine this with a US county shape file to see geographical variations and conduct further analysis. There are two research questions I would like to answer. Firstly, how does obesity vary geographically at a county level? Secondly, can geographical variations in obesity be explained by health, socio-economic, environmental and/or demographic factors?

Tasks and Approach

The analytical tasks used to answer each research question are described below, including which techniques are visual and which are computational. The tools used for the analysis are Python for computation and a combination of Tableau and Python for visualisation.

To understand how obesity varies geographically and to find out if variations in obesity can be explained by health, socio-economic, environmental and/or demographic factors, I will perform the following tasks.

1. Initial Analysis
    a. **Visual**: Create a choropleth using a diverging colour scheme to visualise obesity by county to identify spatial patterns.
    b. **Computational**: Compute summary statistics on obesity. Use the mean and standard deviation to identify outliers and filter the data for outliers. Compare these groups on the map and view the raw data to see if there are obvious differences to help with feature selection. Compare data with high and low obesity by filtering to show counties in the first and fourth quartiles.
    c. **Visual**: Create a map trying different Tableau attributes as the data layer while filtering for high and low obesity. Compare the high and low obesity maps.
2. Data Preparation

   a. **Computation**: Use previous research and initial analysis to define which health, socio-economic, environmental and/or demographic features to include from the original dataset. Check the summary statistics, skewness and missing values for all features to determine data transformation needed.
   b. **Visual:** Plot scatterplots and histograms of the features to check for outliers and normality.
3. Data Transformation
   a. **Computational:** Impute missing columns. Transform data so it looks more normal and the effect of outliers is reduced. This is so the correlation matrix and linear regression are more accurate.
   b. **Visual:** Check scatterplots and histograms again after transformations to confirm.
4. Feature Selection
   a. **Computational**: Generate a correlation matrix of (transformed) features and the target. Use the Pearson correlation coefficient shown in the matrix to establish the linear relationship between each feature and the target [10].
   b. **Visual**: Create a heat map of the correlation matrix to easily identify which features are highly correlated with each other; confirm by checking values in the correlation matrix.
5. Model Build
   a. **Computational**: Create a linear regression model for each feature with moderate to very strong correlation to obesity, i.e. greater than 0.3 or less than -0.3, with obesity as the target [10]. This will be used to analyse each feature individually with obesity and how it varies spatially.
   b. **Computational**: Select the highest correlated features and build a multiple regression model representing all counties for comparison to the single-feature linear models.
   c. **Visual**: Produce scatter plots of each correlated feature including the lines from the linear regression models to check the fit of the line against the data points [7]. This presents the actual values plotted with the best fit line from the linear regression model showing how close the actual and predicted values are to one another for each feature separately.
6. Model Evaluation
   a. **Computational**: Compute the residuals, i.e. the difference between the actual and predicted values, for each linear regression model including the multiple regression model.
   b. **Visual**: For each feature model and the multiple regression model, create choropleths of the model residuals to see if there are geographical differences between the actual and predicted values using that feature. Compare the single-feature choropleths to each other and to the multiple regression model choropleth. This can help explain the reasons for the geographical differences.

Analytical Steps

This section describes the analytical tasks and shows examples for each task. I first look at how obesity varies geographically.

1. Initial Analysis

The choropleth in Figure 1 provides a spatial view of obesity in the United States. There are differences across state-lines, e.g. Texas and Oklahoma, Colorado and Kansas, even though the figures are at county-level. Illinois has mostly low obesity except around the edges where it's closer to states with higher obesity (Figure 2a). Most of the high obesity counties are in the south, but a few are in South Dakota. Comparing the Tableau map to google maps, these areas are Indian reservations. This pattern also occurs in Arizona and New Mexico on another Indian reservation [4] (Figure 2b).

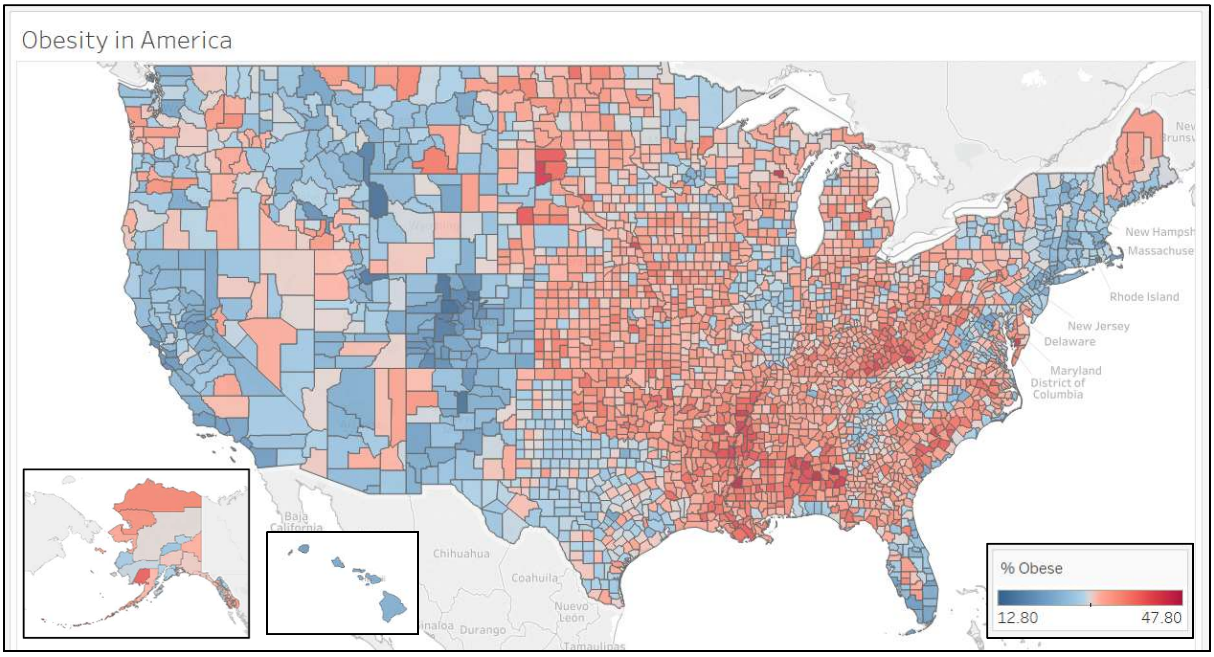

Figure 1. Obesity in America

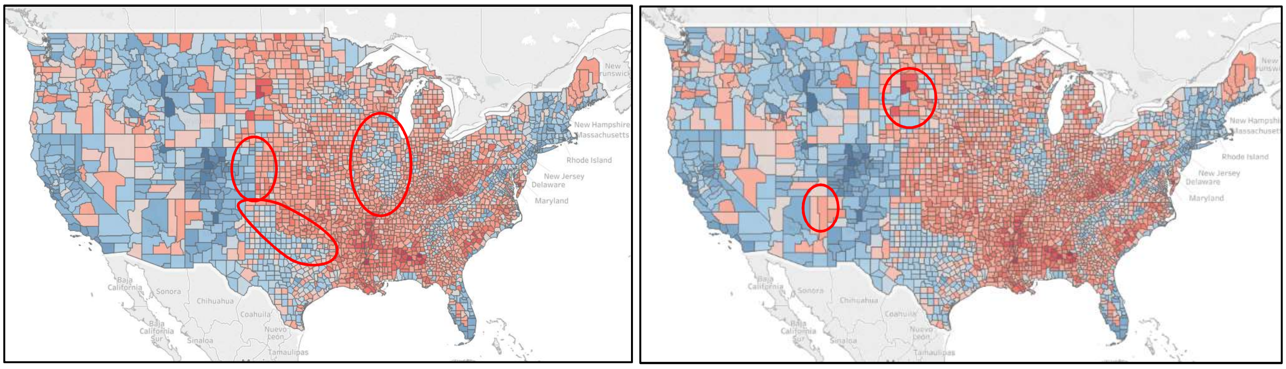

Figure 2. a) Areas of contrast along state lines b) high obesity in Indian reservations

Some of the highest levels of obesity are in the deep south, e.g. Mississippi and Alabama. The 'Southern pattern' of food, including fried food, added fats and processed meats, has been found to result in higher instances of Coronary Heart Disease [9]. This may also explain the high obesity found in the southern states (Figure 3a).

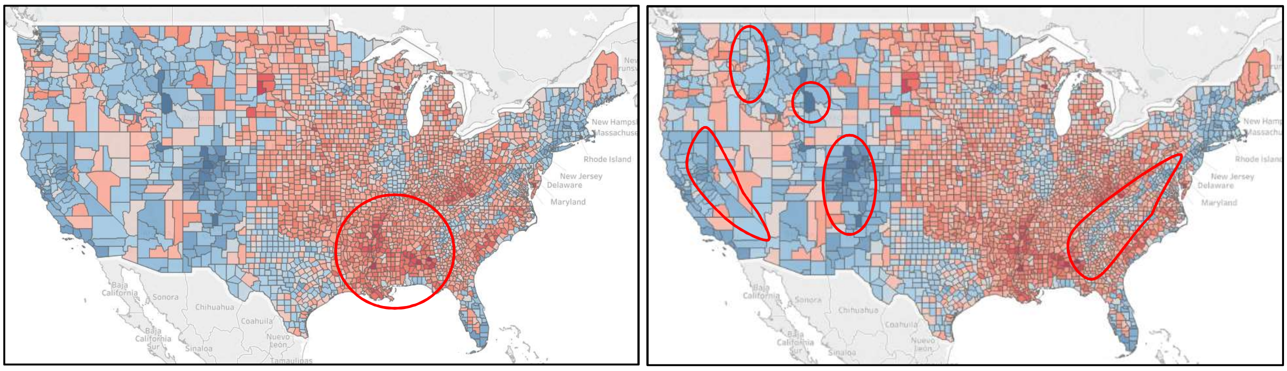

Figure 3. a) High obesity in the deep south b) low obesity around national parks

Some of the lowest levels of obesity are in Colorado. There are patterns of low obesity around national forest and park areas in Virginia, West Virginia, the Carolinas, Tennessee, Georgia, Colorado, California, New Mexico, Idaho and Wyoming (Figure 3b).

Using summary statistics, I identify outliers as counties with obesity higher and lower than three standard deviations from the mean. I also look at high and low obesity using the first and fourth quartiles of data. When comparing the data for the outlier groups, these features stand out as different: Access to Exercise, Food Environment Index, Children in Poverty and Single-Parent Households.

Next, I create a proportional symbol map trying different Tableau attributes while filtering for high and low obesity. When high obesity is combined with household income, a pattern is shown where states with the lowest household income have the highest obesity, except for New Mexico.

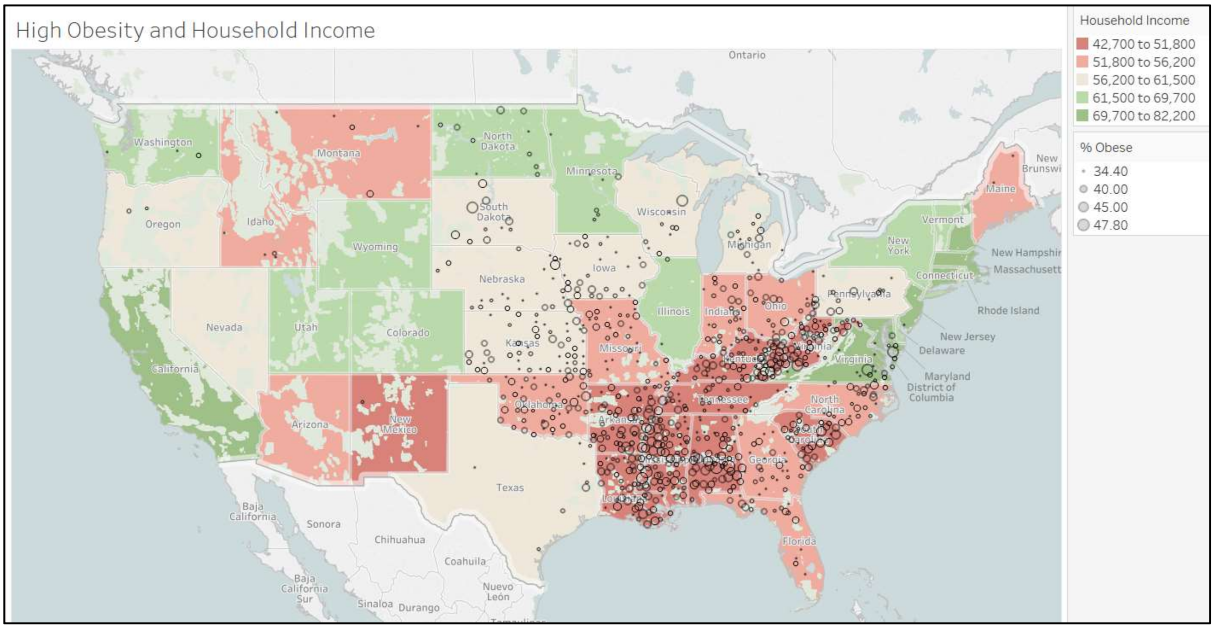

Figure 4. High obesity and household income

2. Data Preparation

Next, I look at whether variations in obesity can be explained by health, socio-economic, environmental or demographic factors. Considering previous research and the above points, I include features related to diet, exercise, education, income, race and gender for the analysis. The dataset does not provide details on areas of nature, but it does contain air quality, so I include this.

After loading the data, I check summary statistics, skewness and missing values for all features to determine the data transformation needed. Then, I create scatterplots and histograms to check for outliers and normality.

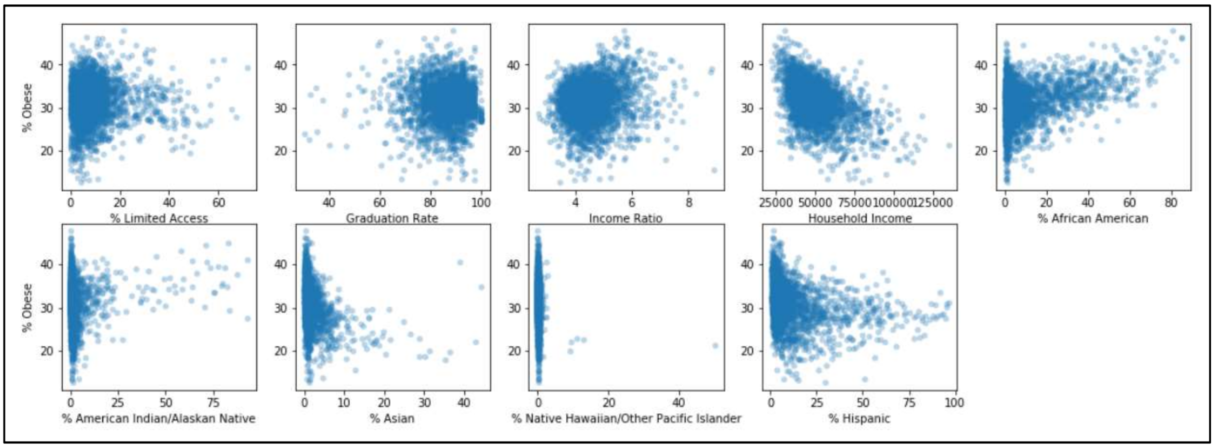

Figure 5. Scatterplots of features with high skew

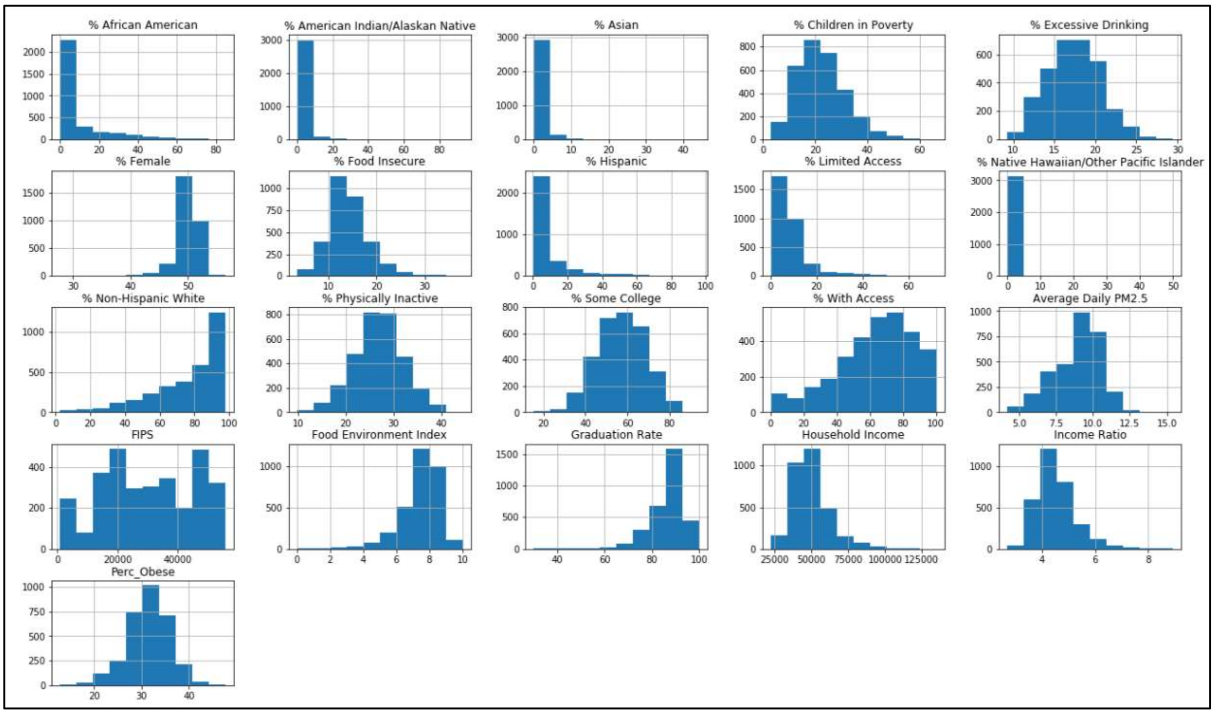

Figure 6. Histograms of all features

3. Data Transformation

I impute missing columns and transform skewed columns so they look more normal. I try different transformations, i.e. log, square root, and inverse, for features using the histograms to evaluate.

I check the scatterplots and histograms to see the transformation. This also reduces the effect of outliers. Most of them were able to be transformed, but a few did not change and were left as is.

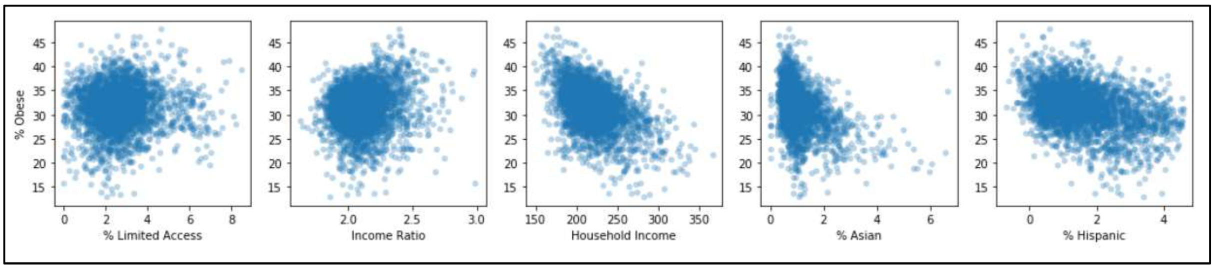

Figure 7. Scatterplots after data transformation

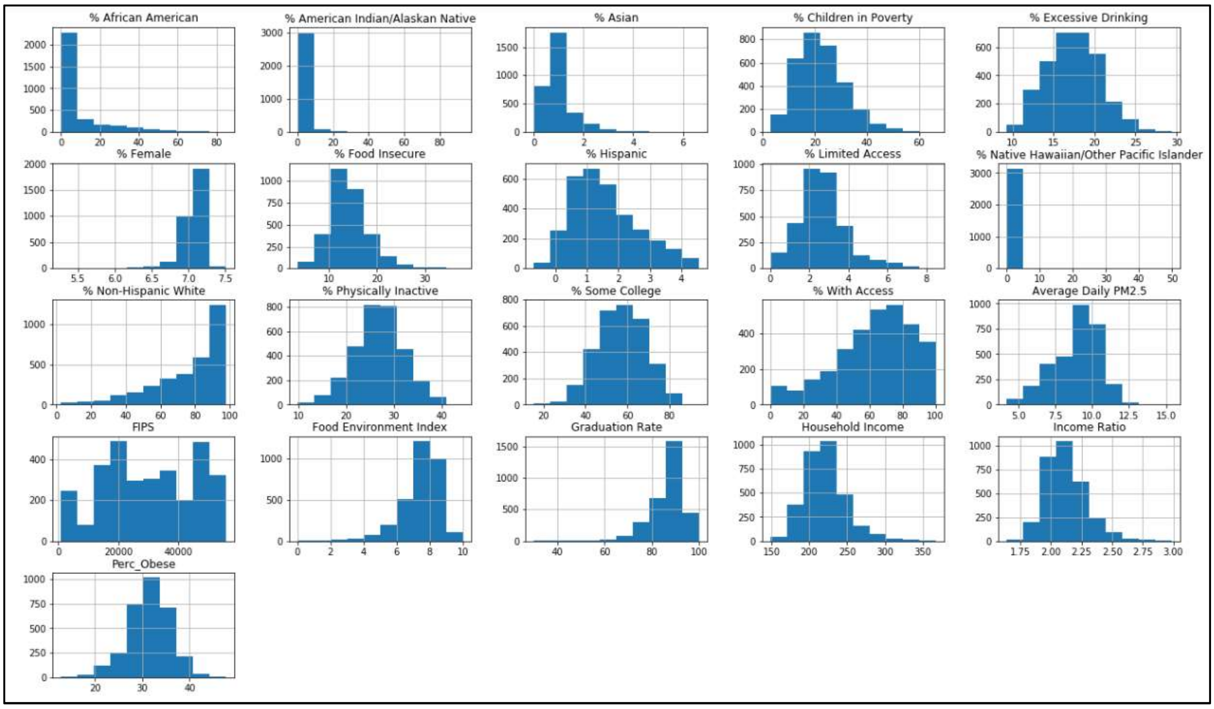

Figure 8. Histograms of features after data transformation

4. Feature Selection

Next, I generate a correlation matrix of the features and the target to identify which features relate highest to obesity. I then create a heat map of the correlation matrix to easily see which features are highly correlated with each other. I confirm by checking values in the correlation matrix.

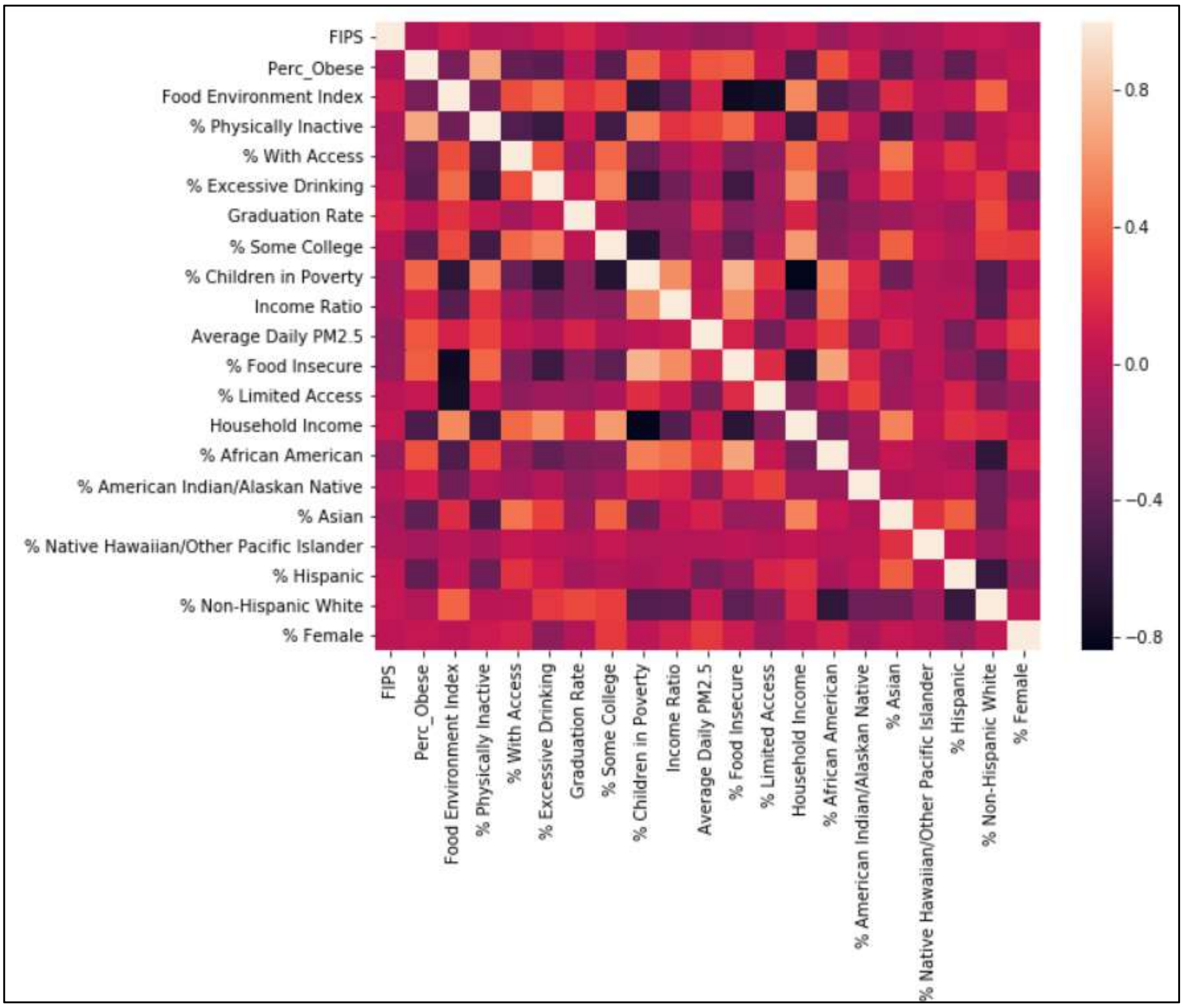

Figure 9. Heatmap of the correlation matrix

5. Model Build

I now create a linear regression model for each correlated feature with obesity as the target. I also select the highest correlated features (plus % American Indian/Alaskan Native to consider Indian reservations) and build a multiple regression model for comparison to the single-feature linear models. The multiple regression model has an R-squared of 0.99, which means it explains 99% of the variation [10]. It has an average error of 8.13%.

I produce scatter plots of each feature and model to check the fit of the line against the data points.

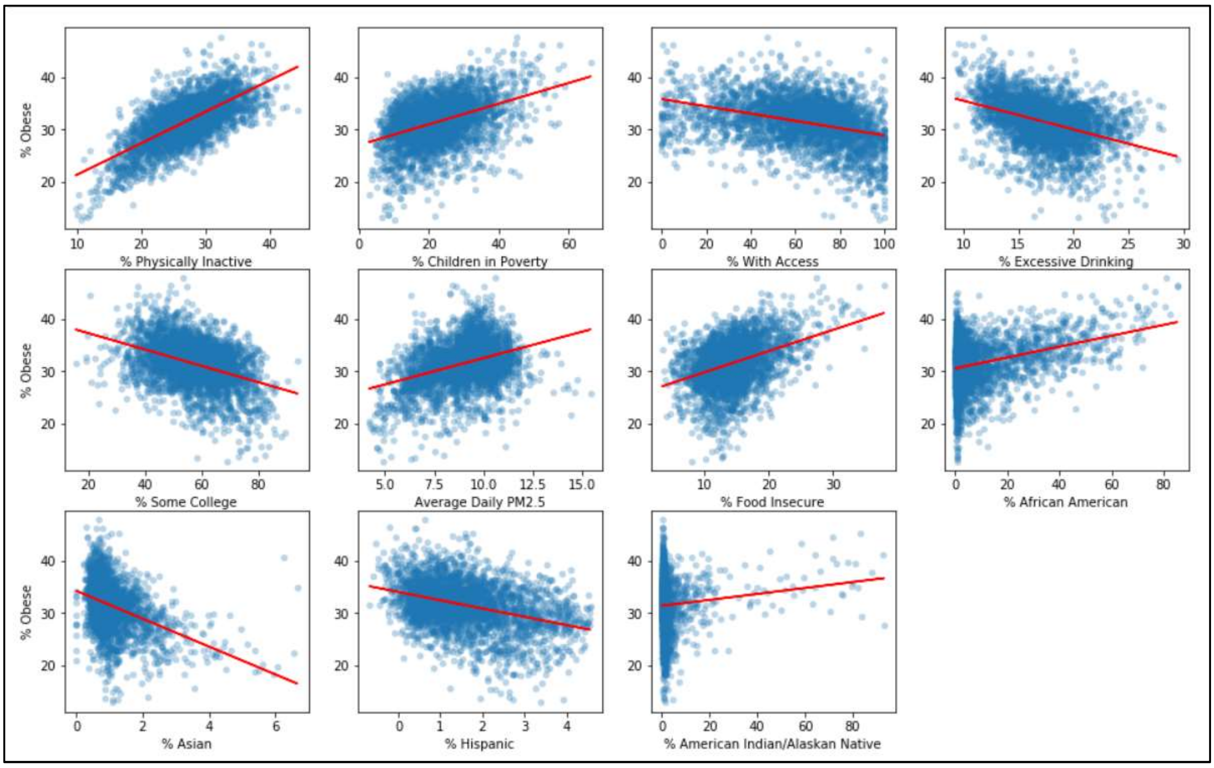

Figure 10. Scatter plots with regression lines

6. Model Evaluation

After confirming the lines are a good fit, I compute the residuals for each model including the multiple regression model. I create choropleths of the residuals to compare the obesity variation by feature.

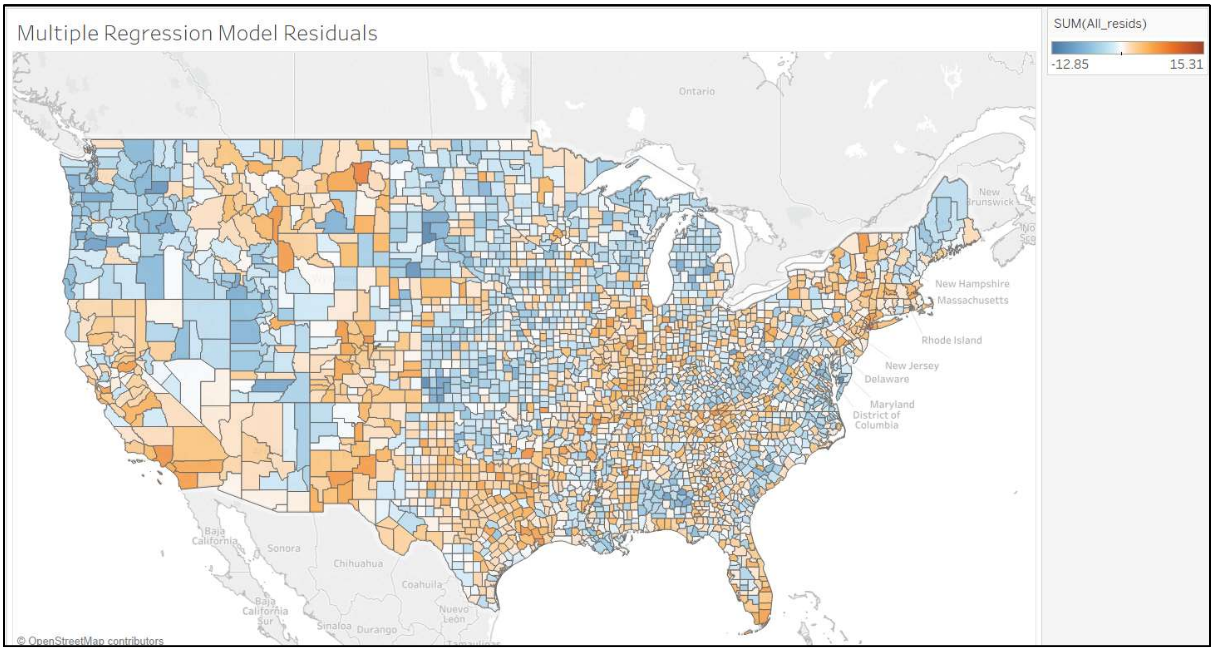

Figure 11. Choropleth of residuals for multiple regression model

The residual maps help explain some of the initial observations. The counties in Kansas that border Colorado are best modelled by Excessive Drinking (Figure 12b), Physically Inactive (Figure 12a) and % Asian (Figure 13a). The Colorado counties next to Kansas are best modelled by Physically

Inactive (Figure 12a) and Air Quality (Figure 13b) when looking at single features, but they are best modelled by the multiple regression model (Figure 11).

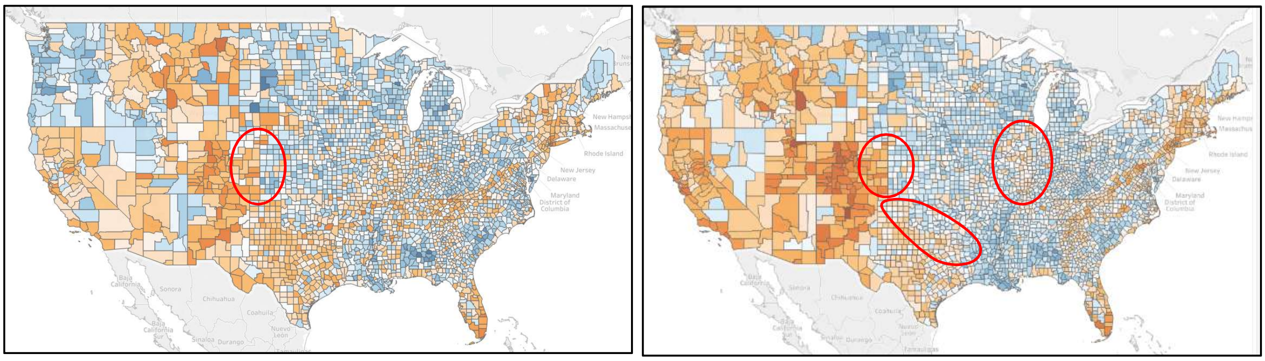

Figure 12. a) Residuals for Physically Inactive b) Residuals for Excessive Drinking

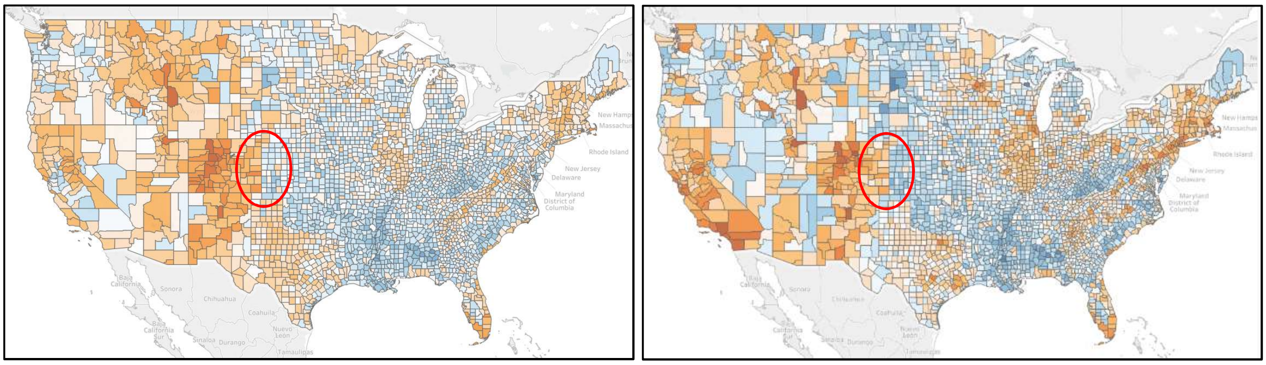

Figure 13. a) Residuals for % Asian b) Residuals for Air Quality

The differences between Texas and Oklahoma and Illinois' low obesity compared to its nearby states are best explained by Excessive Drinking (Figure 12b).

The areas with Indian reservations don't follow a consistent pattern in any model, so I cannot explain the high obesity found in those areas. Even the model using % American Indian/Alaskan Native did not predict obesity well in these areas.

High obesity in the deep south is not represented well by any of the models. The best representation is by Food Insecurity, but there is still a lot underestimated by that model.

Every model overestimated obesity in Colorado. The best model is the multiple regression model.

Obesity around national forest areas are usually overpredicted by the models. Again, they are best modelled by the multiple regression model.

Findings

The residuals of each model helped to explain some of the geographic variation in obesity. Considering the research questions again, I've summarised my key findings from the analysis below.

1. How does obesity vary geographically at a county level?
2. Can geographical variations in obesity be explained by health, socio-economic, environmental and/or demographic factors?

The initial analysis showed that states with the lowest household income had the highest obesity. Some of these were states in the deep south. However, high obesity in the deep south is not represented well by any of the individual models.

Excessive Drinking is the best feature for modelling obesity in the Kansas counties that border Colorado and the counties that border Texas and Oklahoma. It also explains the low obesity in Illinois and the high obesity in surrounding counties.

Physically Inactive is the best feature for modelling the Colorado and Kansas counties that border each other.

Other features that are important are % Asian which models the Kansas border counties well, and Average Daily PM2.5 (air quality) which helps to explain Colorado's border counties.

Some spatial patterns could not be explained well by the residuals. The high obesity in the Indian reservations and the deep south and the low obesity around national forest and parks showed no consistent pattern in the single-feature models. Also, every model overestimated obesity in Colorado. The best model for all these areas is the multiple regression model.

The multiple regression model explains 99% of the obesity variance with 8.13% average error. It includes the following features: % Physically Inactive, % Children in Poverty, % With Access (to exercise), % Excessive Drinking, % Some College, Average Daily PM2.5, % Food Insecure, % African American, % Asian, % Hispanic, and % American Indian/Alaskan Native.

In summary, Excessive Drinking, Physically Inactive, % Asian and Average Daily PM2.5 played the biggest part in spatial variation of the identified patterns. When looking at America as a whole, the health, socio-economic, environmental and demographic features included in the multiple regression model above are important in explaining obesity.

Critical Reflection

This analysis of spatial variation of obesity in America, while providing limited explanations of variation by county, could be used to understand why certain counties have higher obesity compared to other counties. Potentially, those counties could take measures based on these findings to work to reduce obesity.

The visual analytics approaches used were valuable in identifying important features related to obesity. The choropleth used in the initial analysis was beneficial to understand spatial patterns in the data which could not have been done by viewing the data alone. The scatter plots combined with the linear regression lines were helpful to confirm the model fits before proceeding further. The model residuals shown on choropleths were especially useful in understanding how each single-feature model varied spatially, even though it produced limited results. Overall, I was able to partially answer my research questions. I was able to explain variation for some spatial patterns identified, but not all of them.

There are some limitations with the data used. The data is a compilation from different sources for a period of eight years. Some of the columns provided span multiple years and some are for a single year. Obesity figures are from 2014, but some of the other data is from before or after 2014, which means it may be outdated or collected after 2014. It's also at a county level and not at an individual level. So, this analysis cannot explain why certain individuals are obese; it only attempts to explain why obesity is different for different counties. There are also limitations in the individual features provided in the dataset. For instance, there is nothing provided describing areas of national forest or parks which were shown to coincide with low obesity in the initial analysis. Also, some of the demographic features were very highly skewed and none of the transformations were able to change them to a normal distribution. This could have affected the correlation values or the linear models. Lastly, I have considered the possibility that there may be differences in how the data is collected for each state, given the differences seen between obesity across state lines, even though it is by county.

The techniques used in this analysis could be applied to other public health problems or indeed other domains. Spatial data, such as data provided at a census tract, county, or possibly state level, for a given problem could be combined with a shape file for initial visual analysis. From there, the same analysis and modelling techniques could be applied. This could be used to analyse spatial differences in chronic illnesses, crime, or voting outcomes. Depending on the data used, some of the same limitations as above may apply. However, the visual analytics techniques themselves are appropriate for other applications and domains of a spatial nature.

References

[1] Centers for Disease Control and Prevention (2018) *Adult Obesity Facts*. Available at: https://www.cdc.gov/obesity/data/adult.html (Accessed at: 12 December 2018).

[2] County Health Rankings & Roadmaps (2018) *Rankings Data and Documentation*. Available at: http://www.countyhealthrankings.org/explore-health-rankings/rankings-data-documentation (Accessed: 30 October 2018).

[3] Curioni, C., Lourenco, P. (2005) 'Long-term weight loss after diet and exercise: a systematic review', *International Journal of Obesity*, Volume 29, pp. 1168-1174. doi: https://doi.org/10.1038/sj.ijo.0803015

[4] Google Maps (2018) Available at: https://www.google.com/maps (Accessed date: 12 December 2018).

[5] Hales, C., Carroll, M., Fryar, C., Ogden, C. (2017) *Prevalence of obesity among adults and youth: United States, 2015–2016*. NCHS data brief, no 288. Hyattsville, MD: National Center for Health Statistics. Available at: https://www.cdc.gov/nchs/data/databriefs/db288.pdf (Accessed: 12 December 2018).

[6] Harrington, D., Elliott, S. (2009) 'Weighing the importance of neighbourhood: A multilevel exploration of the determinants of overweight and obesity', *Social Science & Medicine*, Volume 68, Issue 4, pp 593-600. doi: https://doi.org/10.1016/j.socscimed.2008.11.021.

[7] Keim, D. et al. (2010) *Mastering the Information Age Solving Problems with Visual Analytics.* Edited by Daniel Keim et al. Gosler: The Eurographics Association.

[8] Robert, A., Reither, E., (2004) 'A multilevel analysis of race, community disadvantage, and body mass index among adults in the US', *Social Science & Medicine*, Volume 59, Issue 12, pp 2421-2434. doi: https://doi.org/10.1016/j.socscimed.2004.03.034

[9] Shikany, J., Safford, M., Newby, P., Durant, R., Brown, T., Judd, S. (2015) 'Southern Dietary Pattern Is Associated With Hazard of Acute Coronary Heart Disease in the Reasons for Geographic and Racial Differences in Stroke (REGARDS) Study', *Circulation*, Volume 132, No. 9, pp. 804-814, doi: 10.1161/CIRCULATIONAHA.114.014421.

[10] Turkay, C. (2018) 'Week 04 Investigate relations (& structures)' [PowerPoint Presentation]. *INM430: Principles of Data Science*. Available at: https://moodle.city.ac.uk/pluginfile.php/1527026/mod_resource/content/0/INM430_Week04_Lecture.pdf (Accessed: 17 October 2018)