Background

For this project I chose to work with a few different datasets. The primary two datasets I used are the locations of every McDonald's restaurant in the United States, located at https://introcs.cs.princeton.edu/java/data under mcdonalds.csv, and a table from https://www.cdc.gov/obesity/data/prevalence-maps.html containing each state's obesity rate in 2017. The McDonald's dataset contains the latitude, longitude, name, and address for the nearly 14,000 McDonald's locations in the United States. The address field is a bit messy, and contains the street address, city, two letter state abbreviation, and telephone number. The obesity table contains just two variables: the state (fully spelled out) and the obesity rate in 2017 for each of the 50 states plus Washington DC. Because I planned to merge the two tables, I also downloaded a dataset that contains each of the 50 states plus their two-letter abbreviation. For the project, I am interested in seeing which cities and states have the most McDonald's locations. In addition, I want to see if there is any relationship between the amount of McDonald's locations and obesity rates. Finally, because I want to account for state's population, I plan to use per-capita data, so I downloaded a dataset with each states' population to be able to account for this. I later added a dataset that maps each state to a region in order to enhance one of my plots.

Part I: Reading in the Data

For each of my datasets, I initially read in character data using wider than needed columns to check if any of the data would be cutoff. I also used the notes within SAS to check if the maximum length of any of the records exceeded any of my column lengths. I also confirmed that everything looked alright using simple proc print and proc freq statements, to make sure there was no unexpected missing data and to get an idea of what the data looks like. Each of the datasets were located on a GitHub repository, which I first had to read in before completing each data step.

Part II: Data Wrangling

For both the McDonald's datasets and the obesity dataset, there was some cleaning and merging necessary to prepare both datasets for further analysis and visualization.

First, in the McDonald's dataset, I needed to extract information from the address field, specifically the state and city of the McDonald's restaurant. Nearly every address had the same format, with each value being separated by commas, thus I used the scan function to extract both the city and state. However, for the state value, there was occasionally some erroneous information included before the next column. To deal with this, I used the substrn function to take the first two letters of the state field, which would ensure that nothing after the state's two letter abbreviation would be included.

Because the McDonald's dataset had each state's abbreviations, while the obesity and population dataset only had the full state name, I added state abbreviations to the two datasets by merging them together with the state abbreviation dataset. I also deleted state values that were not the 50 states plus DC, as those were the only places in the McDonald's dataset.

For one part of the analysis, I planned to map which cities have the most McDonald's locations. However, in New York City, the locations are separated into boroughs (thus the city is 'Brooklyn' etc). To fix this, I created a temporary dataset that would convert the city to New York if it found any of the five boroughs in the McDonald's address field.

I did quite a bit of analysis, some of which required some more data-wrangling, which I will describe alongside the specific analysis in the next section.

Part III: Data Analysis

The first thing I wanted to do was map the ten cities with the most McDonald's in the US. Using the dataset described above, I grouped the data by city and state pair using proc sql and counted how many pairs of each were present in the data, saving it in a new dataset, *cities*. I first tried to use proc freq but it could not handle the amount of city/state pairs. I made sure to use city/state pairs because there can be multiple cities with the same name in the US. After grouping the data, I sorted the dataset descending by the count, and then added the latitude/longitude coordinates for each city by merging the data with the original McDonald's dataset. Finally, I added a 'label' column by concatenating the sorted row number and the city, so the city rank would show on the map. Now the data was ready to be mapped, which I did using sgplot, and can be seen as *Figure 1* in the appendix. Unsurprisingly, McDonald's are most common in some of the biggest cities in the US, as the top three cities are New York, Houston, and Chicago, respectively. I also used proc print to show these full results in table form, including the count for each city, as I was interested to see if there were any cities that stood out.

For the next part of my analysis, I looked at the states with the most McDonald's locations, as I will later compare this each state's obesity rate. In doing so, I first use proc sql to create a new dataset containing the number of each McDonald's in each state. Next, I merged population data into the dataset, and created a new variable accounting for the number of McDonald's in each state per 100,000 people (I used 100,000 people because it created clean single-digit numbers). After sorting the data, I created a bar chart to visualize the results. As you can see in *Figure 2*, the states with the most McDonald's per person are Kentucky, West Virginia, and Michigan.

Next, I wanted to see which states have the highest obesity rate to see if there were any similarities in states at the top of both lists. As seen in *Figure 3*, there seems to be some common states that are high in both McDonald's per capita and obesity rate. However, I wanted to do a take a closer look at this relationship and look at the relationship between these two variables more formally.

For the final part of the analysis, I looked closer at the relationship between obesity rate and McDonald's per capita within each state. To do so, I first merged the two datasets containing each state's McDonald's per capita and each state's obesity rate. Next, I created a scatter plot using sgplot, including a regression line. Looking at the results in *Figure 4,* we see a pretty clear positive relationship between McDonald's per capita and the obesity rate per state. I also added data to color each point by their census region, to better help visualize the regional cluster of the data. Now, you can clearly see that there is a cluster of states in the top right corner composed of

from the Southern region, indicating that it may be the unhealthiest region in the US. There is also a cluster of states from the Midwest in the upper right, while states from the west and northeast appear to be the healthiest. Using proc corr, I found the correlation coefficient between obesity rate and McDonald's per capita, which is .38. I also performed a basic regression analysis, regressing obesity rate on McDonalds per capita to take a closer look at this relationship. I found that there is a statistically significant relationship between the two variables with a beta-coefficient of about 2.0, as seen in *Figure 5,* meaning that an increase of 1 McDonald's per 100,000 people corresponds to a 2.44% increase in obesity rate. Looking at the diagnostics, there appeared to be a few outliers, namely DC and Hawaii. Because of this, I also created a new dataset removing these two observations and reran the regression analysis and again found a significant relationship with a .49 correlation coefficient.

Part IV: Appendix

Figure 1

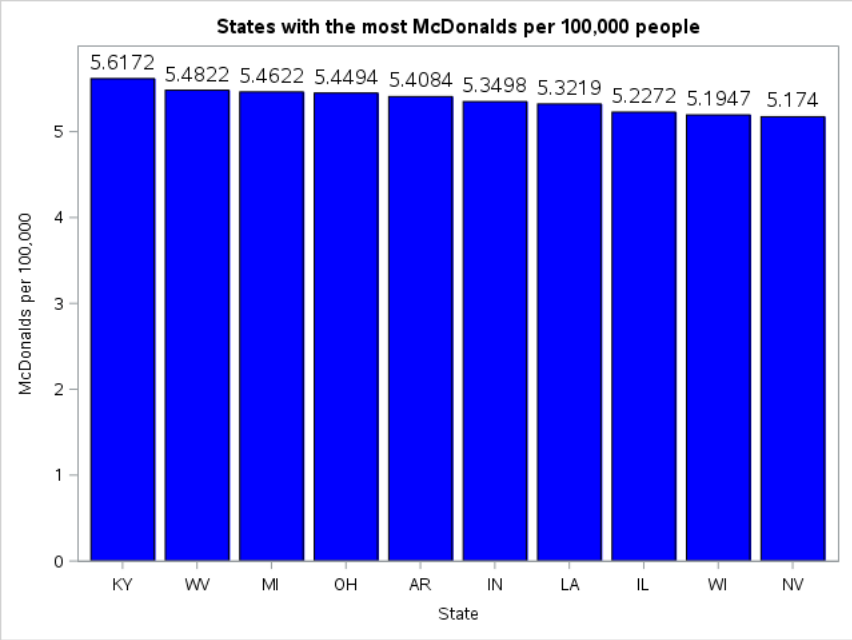

Figure 2

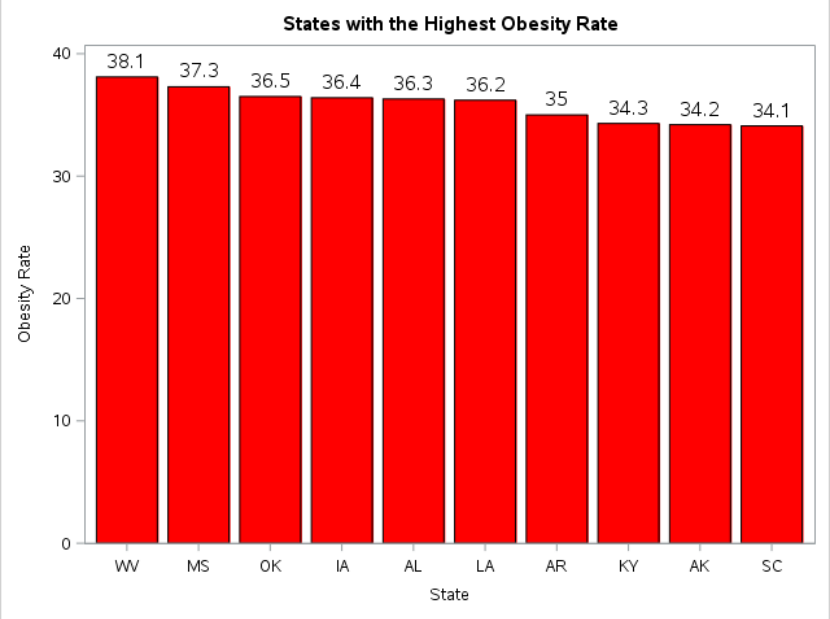

Figure 3

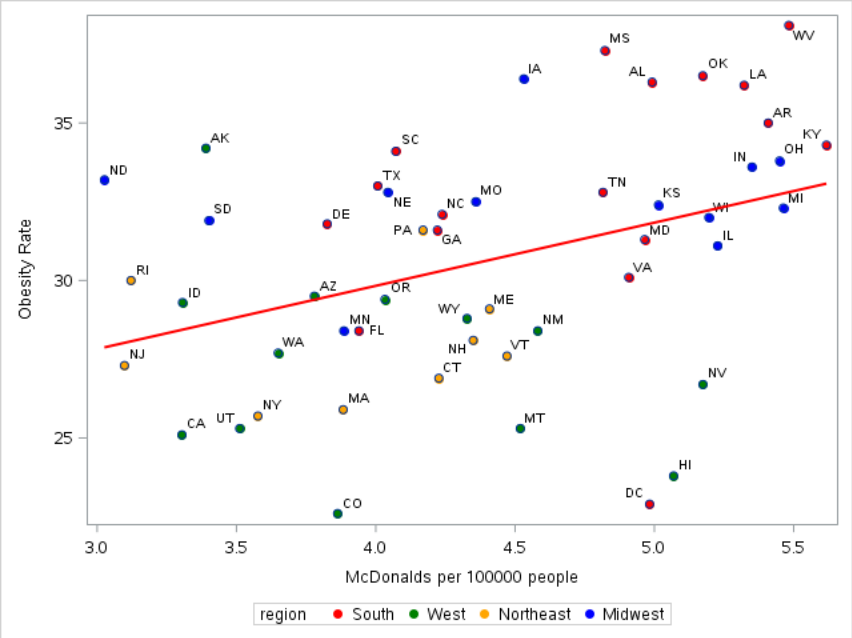

Figure 4

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	20.23790	2.77088	7.30	<.0001
per_100000	McDonalds per 1000000	1	2.44578	0.62731	3.90	0.0003

Figure 5