****

**Nassau Re - Life Insurance Public Data Case Study (Project Report)**

Team Members: Daniel Zygadlo, Sammy Soumakis, Jeanne Wang, Christopher van Schalkwyk, Ze Ming (Ben) Chan, Haoxin Xu

**Executive Summary**

The Nassau Re case study required the team to utilize public data to investigate drivers of mortality differences across the United States and to identify strategies for life insurance company usage. Therefore, we created a dataset with 51 rows and 15 columns, each row representing a state and each column a different variable. Furthermore, RStudio was utilized to create linear models, visualizations and a new variable. The team goal is to create a state adjustment factor, a factor that takes an individual’s environment into account when pricing their premium, for each state. To accomplish this goal, we had to learn more about how the environment within the state could have an effect on the life expectancy of that particular state, and consequently, the behavior of the policyholder. More specifically, certain regions tend to have various characteristics associated with them.

What we discovered from our linear regression and visualizations was that there is a strong association between life expectancy, obesity, smoker status, income and several other factors when analyzing the data by state. Seeing that different factors had similar trends on life expectancy, we created state adjustment factors. State adjustment factors will be beneficial in two main ways. First and foremost, it would benefit the life insurance company and the policyholder financially by lowering premiums and increasing life expectancy for that individual. The second benefit is an easy marketing strategy; market benefits to individuals in unhealthy states as an incentive to become healthier and pay lower premiums for their life insurance.

**Introduction**

Life insurance is an industry that is sensitive to changes in mortality. Life insurance companies make assumptions based on how long they anticipate a policyholder to live for. They then use those assumptions to calculate premiums. Therefore, making the right observations and assumptions are one of the most important aspects of insurance - if done incorrectly, companies could lack funds to pay beneficiaries and consequently become insolvent. To prevent this from happening, insurance companies spend a multitude of resources collecting and analysing data to ensure that mortality assumptions are as accurate as possible.

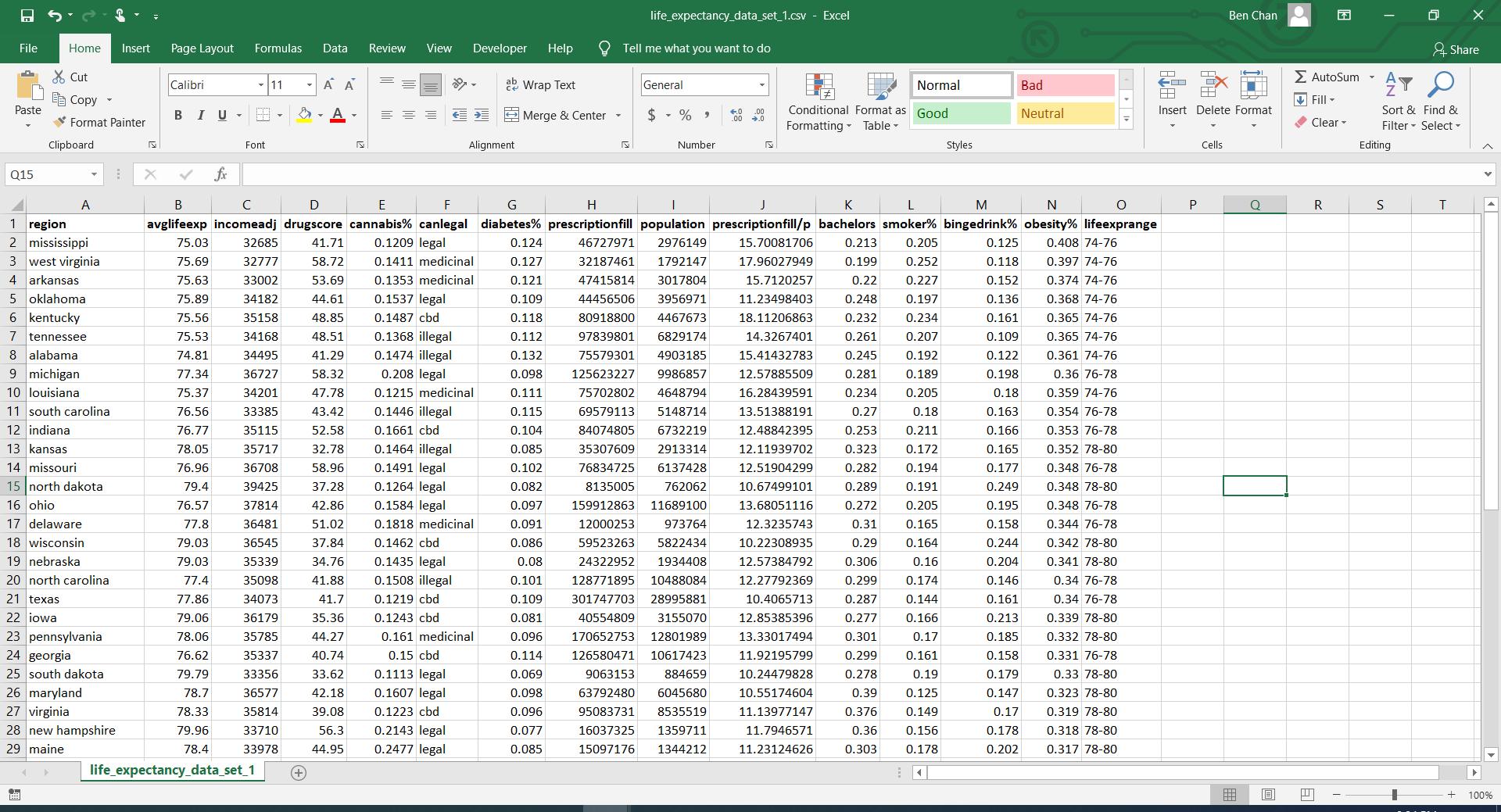
In today’s world of Big Data, we have a plethora of data sources to choose from. Sophisticated data collection techniques such as telematics can bring numerous benefits to an insurer such as making it easier to monitor consumer behavior. However, it may be more cost-efficient for insurers to rely on outsourcing data from parties specialized in data collection so resources can be maximised. In particular, insurers should consider the use of publicly available datasets, which are both inexpensive and easily accessible.

The purpose of this project is twofold. The first is to obtain public mortality data and analyze factors that cause mortality differences. Mortality factors that we had considered included obesity, drug use, diagnosis of diabetes, etc. Secondly, we will use the results of our analysis to develop strategies for Nassau Re to better market its life insurance products. Overall, the project is meant to help improve mortality analysis regarding factors that may otherwise not have been looked into before. This way, Nassau Re will be able to gain deeper insights into the nature of its policyholders.

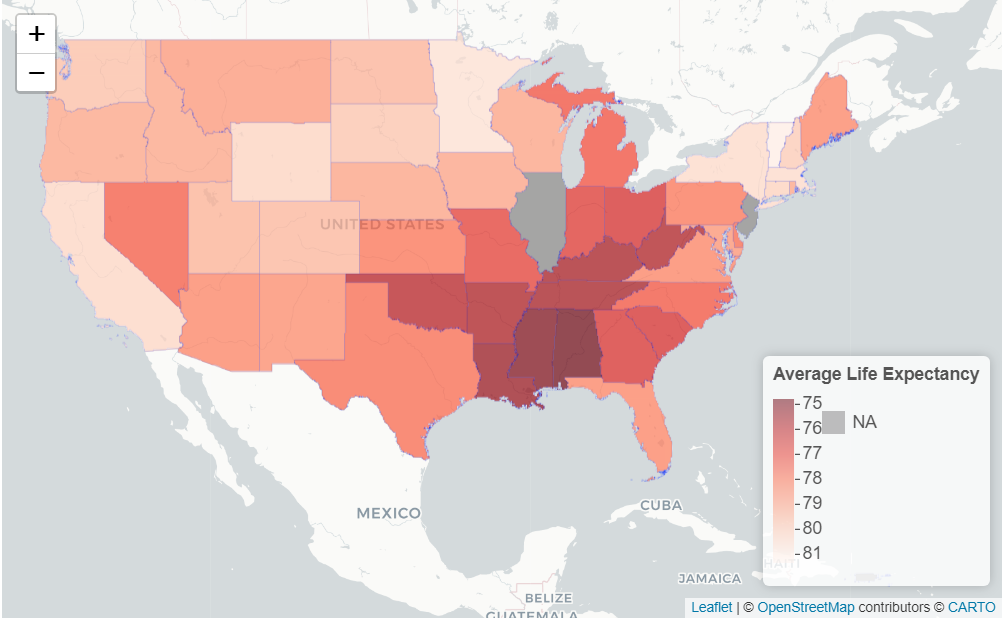
The findings of the project include a connection between mortality and obesity by state as well as recommended state adjustment factors.

**Data and Model**

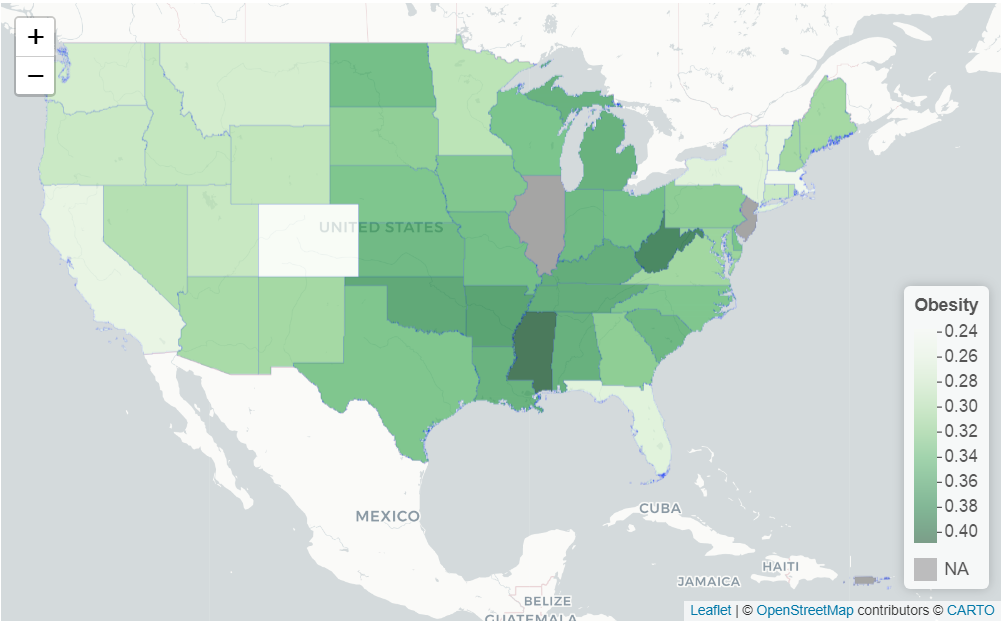
Our initial assumption was that the average life expectancy of US citizens would primarily differ across states. We considered numerous mortality drivers and variables unique to each state and chose the ones that were potentially more significant to include in our model. These include adjusted income (for COL), drug usage, cannabis usage, legality of cannabis, percentage of diabetics, population size, number of retail prescriptions filled, percentage who are highly educated (achieved a bachelor's degree or higher), percentage of smokers, and percentage of binge drinkers. We collected datasets from reliable sources such as the US Census Bureau and the CDC for each mortality factor, and compiled them together into an all-inclusive database in Excel. A snippet of the dataset is shown as follows:



R is the primary program used in this case study to process the data. Not only is it versatile in production of visualization creation, it allows for convenient adjustments and modelling of data. The packages contained in the library are also very helpful for model creation and analysis. Some of the main packages we used include caret, ggplot2, and leaflet - a comprehensive list can be found in Appendix B. The leaflet package, in particular, was very useful as it allowed us to illustrate differences of the mortality factors between each state on an interactive US map. After installing the various packages, factors were assigned to the variables. Geographical codes from US cities and states were then aggregated to create spatial data, and the U.S. territories were removed from the dataset. We then ordered the data by regions, combined the dataframes, and used the leaflet package to create a visualization for average life expectancy, as shown below:



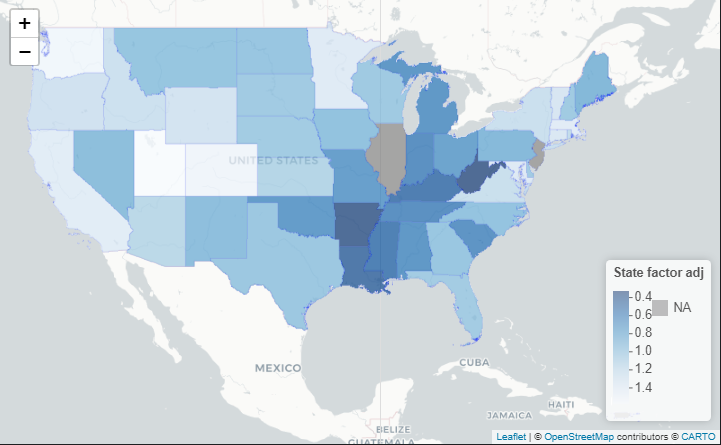
At first glance, it can be seen from the concentrated red in the lower middle right that the states in the deep South seem to have a lower average life expectancy, with Alabama having the lowest at 74.81 years. Then, to observe if our selected mortality drivers followed the same trend, we drew up another visualization for obesity rates:



Interestingly, a clear contrast could be seen again between the South and the other regions (West, Northwest, and the Midwest). For example, Mississippi has a 40.8% obesity rate while Colorado has a 23.8% obesity rate. As we generated visualizations for each other mortality factor, the general trend seems to be followed: states in the deep South have a larger accumulation of negative factors and in consequence, lower average life expectancy. Thus, we can assume the state a person lives in has a significant effect on their life expectancy and we can create state-adjusted factors. These factors indicate how long someone from a specified state is expected to live.

To differentiate the p-values, we took the variables from the linear model from earlier. We wanted a way to come up with adjustment factors. To do so, we took p-values from linear models as weights. For each individual variable we took (1 - p) and made it the weight for each variable. All the p-values were then summed together. This new value was applied to each variable in order to make a weighted average out of it, which is the denominator. Something we noticed was that average prescription orders had an outlier p-value, so we decided to make it more significant. Our reasoning here was that the prescription variable was likely getting drowned out by the other variables. For each variable, we ranked the variables from 1 to 50, with 1 being the worst and 50 being best - the higher the number the healthier the state. Then, each variable rank was multiplied with the p-value weight to obtain the state-adjusted factors.

**Analysis**



A map visualization of the state adjustment factors is shown above. Again, the trend is obeyed, with Southern states exhibiting the lower state-adjusted factors. These factors were created by utilizing an average of every states’ variable ranking weighted according to their p-values from earlier analysis. By taking each states’ ranking and subtracting from the mean we generate an estimate of the deviance average health of the state. This allows us to determine which state is healthy and unhealthy in relation to the average. A lower state-adjusted factor signifies a less healthy state on average.

To quantify the accuracy of the factors, a linear model was created using the adjustment factors as our predictor variable, and life expectancy as our response variable. The model was able to achieve high significance with the state adjustment factor having a p-value of 6.98e-13 and adjusted R-squared of 64.71%. However, these values are likely to be biased as p-values were being used, causing already known mortality indicators such as obesity and smoking frequency to heavily weight each states’ adjustment factors. Regardless the model shows how our new, not as known variables also play a role in predicting mortality trends across the US.

Based on each adjustment factor we are able to market our policies in different ways. For example, in healthier states, we are able to advertise lower premiums or larger future benefits. We can differentiate healthy and unhealthy states based on these adjustment factors.

**Conclusion and Recommendation**

By observing the results displayed by the data and the model, we found that the southern part has the darkest color. This means that people in the South are relatively more unhealthy compared to other regions. This is an obviously huge market for insurance companies. There is a preferability for healthier policyholders in this region as they are able to distinguish their health from what their environments would indicate. So, we suggest designing some relative insurance or insurance policies based on the results of this research to realize the profitability of insurance companies and make some changes to people's living habits.

Through the model, we can distinguish different policyholders well. For different types of policyholders, we recommend setting different premiums and policies for them. First of all, we recommend that insurance companies increase insurance premiums in these relatively unhealthy areas for two reasons. According to the obesity map, we found that people in the South region have more serious obesity. Compared with people with a healthy weight, obese people are at increased risk of many serious diseases and health conditions. (CDC, Obesity) This means that people in the southern region are more likely to get high blood pressure, diabetes, and other cardiovascular diseases. Their bodies are more prone to health conditions than people in other states. What’s more, because the risk in the southern region is higher than that in other regions, insurance companies have to increase their claim reserves.

Also, in order to better improve people's living habits, we recommend offering a financial incentive. We can appropriately lower premiums for healthy people in relatively unhealthy areas. We found that the environment can affect people's living habits. So when a person is surrounded by people who adhere to a healthy schedule and have healthy eating habits, they will be affected in a subtle way, begin to change, and slowly become healthier. (Stulberg, B) Lowering premiums for healthy people creates an incentive for ‘riskier’ individuals to adopt positive lifestyle changes so they will be awarded lower premiums as well.

Insurance companies can use this to combine different datasets to design suitable insurances for different types of insurance. They can rely on the results of data visualization to enrich the types of insurance and refine different policyholders. More personalized insurance clauses will be applicable to more different classes and be more relevant to people's lives which will expand the customer base. Insurance companies will also be able to better allocate claim reserves, so as to reduce risks and increase the profits of insurance companies.

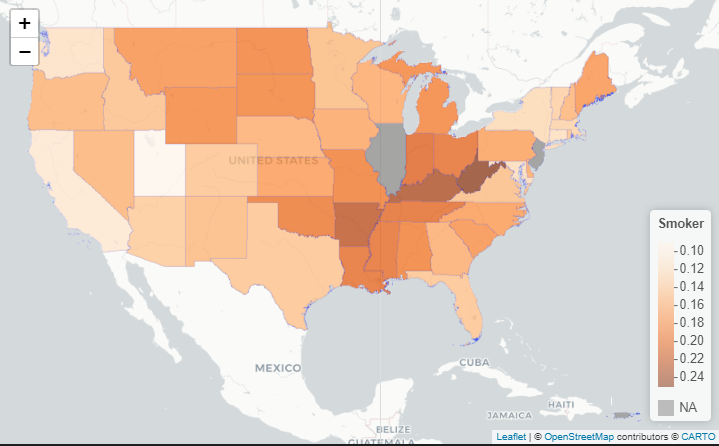
**Appendix**

A. R Packages

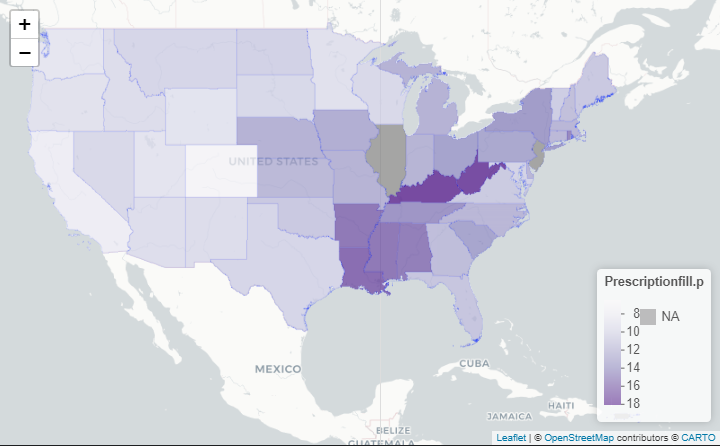
* caret - Meant to make the modeling process more convenient
* e1071 - For training and testing datasets
* rpart - For building regression trees
* rpart.plot - Used to visualize regression tree models
* randomForest - Averages out many regression trees
* mlbench - Brings in machine learning benchmark problems
* gbm - Used for boosted regression trees
* Data.tree - Creates trees for hierarchical clustering
* ggplot2 - Used for creating modeling visuals
* leaflet - Creates interactive maps
* rgdal - Brings map data into R
* dplyr - Allows for data manipulation
* rnaturalearth - Visualizes map data
* WDI - Brings in data from the World Bank
* tigris - Puts lines on a map to show state borders

B. Other Map Visualizations

**Geographic Trends of Smokers**



**Geographic Trends of Prescriptions Filled Per Person**



**Works Cited**

Stulberg, B - Harvard Public Health Review. (Fall 2014). The key to changing individual health behaviors: change the environments that give rise to them. <https://harvardpublichealthreview.org/the-key-to-changing-individual-health-behaviors-change-the-environments-that-give-rise-to-them/>

Centers for Disease and Control and Prevention. (2019, June 21). U.S. Life Expectancy at Birth by State and Census Tract - 2010-2015 <https://healthdata.gov/dataset/us-life-expectancy-birth-state-and-census-tract-2010-2015>

Rasmussen University. Salary by State: Where Can You Really Earn the Most? <https://www.rasmussen.edu/student-experience/college-life/salary-by-state/>

Elflein, J. (2021, January 20). Percentage of U.S. adults that have used cannabis within the past year from 2018 to 2019, by state. <https://www.statista.com/statistics/723822/cannabis-use-within-one-year-us-adults/>

DISA Global Solutions. (2021, February). Map of Marijuana Legality by State. <https://disa.com/map-of-marijuana-legality-by-state>

Centers for Disease and Control and Prevention. Diagnosed diabetes prevalence among the US population.

<https://gis.cdc.gov/grasp/diabetes/DiabetesAtlas.html#>

Kaiser Family Foundation. (2019). Number of Retail Prescription Drugs Filled at Pharmacies by Payer. <https://www.kff.org/health-costs/state-indicator/total-retail-rx-drugs/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D>

U.S. Census Bureau. (2019, December 30). State Population Totals and Components of Change: 2010-2019

<https://www.census.gov/data/tables/time-series/demo/popest/2010s-state-total.html>

Centers for Disease and Control and Prevention. (2020, September 14). Map of Current Cigarette Use Among Adults.

<https://www.cdc.gov/statesystem/cigaretteuseadult.html>

Centers for Disease and Control Prevention. (2020, September 30). Data on Excessive Drinking. <https://www.cdc.gov/alcohol/data-stats.htm>

Pareto Software, LLC. (2021, January 31). US Cities Database. <https://simplemaps.com/data/us-cities>

Centers for Disease and Control and Prevention. (2020, September 17). Adult Obesity Causes & Consequences.

<https://www.cdc.gov/obesity/adult/causes.html#:~:text=Obesity%20is%20serious%20because%20it,and%20some%20types%20of%20cancer>.