|  |  |
| --- | --- |
| **LYDIA TSANG Examine Correlation between lifestyle factors on health in Nashville/Tennessee**  A look at State and city level health data for health trend and social factors that determine health | |
| **Executive Summary** | Tennessee has one of the highest chronic disease rates among states. 90 % of health happens outside doctor offices and hospital. [[1]](#footnote-1) Lifestyle choices are one of the major determinants to health.  I would look at chronic diseases indicators (CDI) to examine TN health at state level and look for trends and patterns. Later, I would look at city level and examine Nashville regarding unhealthy behavior such as nutrition intakes preference and physical activities amount.  The desired project outcome is to provide trends and patterns of lifestyles of TN/Nashville level, raise public health awareness and thus improve preventive care on lifestyle changes perspective.  Tools: Python and Tableau  **MVP:** Assess TN health status among states like what is the ranking and highlight demographic groups that have amongst the highest diabetes rates, lowest nutrition intake and physical activities. Use maps for TN and bar charts for showing demographic groups.  Create map with census tract showing communities in Nashville with high concentration of obesity and low physical activities. |
| **Motivation** | My motivation is that healthcare organization strives their best to identify the best ways to deliver high quality care for patients. While it is important to cure patients with the best healthcare services, I would like to contribute efforts to the importance of preventive care. Especially, good nutrition, physical activity, and a healthy body weight are essential parts of a person’s overall health and well-being. [[2]](#footnote-2) . Today, approximately 1 in 3 adults (34.0%) and 1 in 6 children and adolescents (16.2%) are obese. But it is often that people would ignore it and not knowing the seriousness until the potential lifestyle would lead to serious health issues such as obesity and diabetes that need to seek for medical help.  I believe stressing the importance of healthy lifestyle can increase the effectiveness of preventive health on healthcare system. |
| **Data Questions** | Context: There is a close relationship between lifestyle and leading chronic disease such as obese/diabetes. CDI is used to examine health at state level. As a matter of fact, physical activity and nutrition are part of CDI that important aspects to examine. Thus, after knowing the big picture at state level, I would look at lifestyle choices of city level mainly on nutrition intakes and physical activities.  Hypothesis: Diabetes/Obesity is positively correlated to higher CDI indicator of lifestyle(physical activities and nutrition intakes).  Data:   1. File name: U.S.\_Chronic\_Disease\_Indicators\_\_CDI\_   Site: <https://chronicdata.cdc.gov/Chronic-Disease-Indicators/U-S-Chronic-Disease-Indicators-CDI-/g4ie-h725> (data dictionary is also included in the link)  Detail: is a file about a survey of state level from 2010-2018 of answering questions of CDI  Question:  -Comparing other states, obesity and what is highest indictor?  -Clean the data in python  -use tableau to visualize the data on map to explore  2. ~~File name: DiabetesAtlasCountyData~~  ~~Detail: this file tells you state broken down by counties to see which county among states has highest diagnosed diabetes for adults aged 20+ in 2017~~  ~~Question:~~  ~~-where TN is at among other states~~  ~~-In TN, how Davidson county is doing compared to other county~~  ~~-there is no geospatial, use bar chart to answer those questions~~   1. File name: ~~Behavioral\_Risk\_Factor\_Surveillance\_System\_\_BRFSS\_\_Prevalence\_Data\_\_2011\_to\_present\_~~   ~~Nutrition\_\_Physical\_Activity\_\_and\_Obesity\_-\_Behavioral\_Risk\_Factor\_Surveillance\_System~~  Detail: state level, it is a file from 2011-2019 to check different CDI and are broken down by different social descriptors. (age,education,gender,income,race)  -2a. comparing other states, where TN with adult > 18years old with obese , insufficient fruit and veg intake, no physical activity as indicators for ranking the states  -2b who are theses people? Find out their background by looking at TN social descriptors. (obese breakdown by social descriptor in file 1: : U.S.\_Chronic\_Disease\_Indicators\_\_CDI\_.copy)  -across all states, which social descriptor has the highest level (use time series)  -In TN, which social descriptor has the highest level (use time series)  -social descriptors are categorical so how to combine with time series?   1. File: 500\_Cities\_\_Obesity\_among\_adults\_aged\_\_\_18\_years   Detail: it shows in 2017 the obese level among cities across states for adult (18 years or above?)  Question:  -the file has census tracts so we can examine which area in Nashville has more obese concentration  -use geospatial python to map the data  5. File: 500\_Cities\_\_No\_leisure-time\_physical\_activity\_among\_adults\_aged\_\_\_18\_years  Detail: for 2017, the file has census tracts so we can examine which area in Nashville has more concentration with no exercise and see if map together with 4 or compare side by side to check correlation (18 years or above?)  -use geospatial python to map the data  6. city level nutrition intakes?? |
| **Schedule (thru 12/19):** | Week of 11/2: Data Q. 1&2 (2 hours +)  Week of 11/9: Data Q. 3 (2 hours +)  Week of 11/16: Data Q.4 (3-4 hours)  Week of 11/23: Data Q.5 and summarize findings  Week of 12/1: Look for census tract data for Nashville about Nutrition intakes and possibly more up to date year for census tract data (obese, physical activities, nutrition intakes)  12/7: Use dataset found and combine with project  12/8: Data Visualization and think of story-telling plots  12/12: Data Visualization and think of story-telling plots |
| **Data Sources** | Health data comes from surveys, medical procedures, and records reviews. Some data sources are the Behavioral Risk Factor Survey which comes from telephone surveys or birth statistics which comes from the information submitted for birth certificates. Data is available for a number of useful metrics including births, injuries, deaths, potential life lost and more. |
| **Known Issues & Challenges** | Known Issues: I have state level to compare states for obesity, physical activities and nutrition and also the time is up to date from 2011-2019. However, when it comes to narrow down to Nashville, the data is for 2017 only and not sure if that is counted up to date. I like that it has census tract to map. I am still lacking 500 cities for nutrition intakes for Nashville.  Challenge: find more data set in case my dataset findings do not work out.  -Still need to see which aspect stands out more in terms of diabetes/obesity  -not sure if the geospatial files would be consistent to show desired effect  Feedback (Will)  -Narrow down enough to tell the story.  -start data cleaning  -keep in mind MVP |

Reference:

<https://www.tn.gov/health/cedep/environmental/healthy-places/healthy-places/tools-and-resources/tr/data-and-indicators.html>

<https://www.cdc.gov/healthyplaces/toolkit/sources_of_health_data.pdf>

<https://www.cdc.gov/cdi/overview.html>

<https://chronicdata.cdc.gov/500-Cities/500-Cities-Obesity-among-adults-aged-18-years/bjvu-3y7d>

Project Log:

11/17/20

Data Question 1:

* Understanding datatype:

<https://seer.cancer.gov/seerstat/tutorials/aarates/step2.html#:~:text=A%20crude%20incidence%20rate%20is,per%20100%2C000%20population%20at%20risk>.

Next we use population data with the counts to calculate crude rates for each of the 19 age groups. The populations must be for the same geographic area, sex, race, and years (the sum of the black female populations for 2000-2004 in the SEER 9 registries). In this step, we are calculating the crude, age-specific rates. A crude incidence rate is the number of new cancers of a specific site/type occurring in a specified population during a year, usually expressed as the number of cancers per 100,000 population at risk. It is calculated using the following formula:

formula for crude rate

I plan to use one datatype and use crude prevalence to compare among states.

* I also see CDC provide guidelines on data visualization with different comparable variables

<https://www.cdc.gov/visionhealth/vehss/help/view-data-locations.html>

* Use brfss for data source comparison

The Behavioral Risk Factor Surveillance System (BRFSS) is the nation’s premier system of health-related telephone surveys that collect state data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services. Established in 1984 with 15 states, BRFSS now collects data in all 50 states as well as the District of Columbia and three U.S. territories. BRFSS completes more than 400,000 adult interviews each year, making it the largest continuously conducted health survey system in the world

* Needs to think of data cleaning on column Q and R for comparing variables

-column Q can select gender, overall and race

-column R is the sub categories of Q above

11/19/20

~~Crude Rates~~

~~Rates allow for fairer comparisons between geographies with different population totals. Crude rates also account for the total burden of a health outcome to a community. This statistic is calculated as the number of events (numerator) divided by the population at risk (denominator). The population at risk is “a term applied to all those whom an event could have happened, whether it did or not.” For many health statistics, the denominator is simply the population total.~~

[~~https://health.mo.gov/data/mica/CDP\_MICA/AARate.html#:~:text=A%20%22standard%22%20population%20distribution%20is,for%20differences%20in%20age%20distributions~~](https://health.mo.gov/data/mica/CDP_MICA/AARate.html#:~:text=A%20%22standard%22%20population%20distribution%20is,for%20differences%20in%20age%20distributions)~~.~~ I crossed out as I would like to change the data value type for crude prevalence to be consistent to measure with other files like the 500 city (11/26/20)

Crude Prevalance is used instead

-Prevalence = The measured or estimated percentage of people -- weighted to population characteristics – with an attribute or disease during a specific year.

<https://nccd.cdc.gov/cdi/rdPage.aspx?rdReport=DPH_CDI.ExploreByTopic&islTopic=RPH&islYear=9999>

<https://www.slideserve.com/elina/assessing-disease-frequency> (also supports definition of crude prevalence as important measure of disease frequency)

Prevalence: is another major measure of disease in the population • information available from surveys, registries, or investigations Number of existing cases of disease in population in time period Prevalence Rate = Population at risk in same time period x 1000

-DataValueType column: I would choose crude prevalance as it is more fair comparison between states according to the definition above

-Question column: I would need to define several conditions to get rows I need

2. Row contains string ‘Meeting aerobic physical activity’ for people who can meet the guideline for exercise routine

3. Row contains string ‘Obesity’ means people who have obese

I want to select TN and see if 1 &2 have anything to do with 3

11/26/20

-I find out the measure of nutrition intake has different filters and DataValueType ‘Median’

- Question column

Below are the conditions needed to be filtered out

-Median daily frequency of fruit consumption among adults aged >= 18 years

- Median daily frequency of vegetable consumption among adults aged >= 18 years

-I decide to use excel to do pivot chart for exploration since there are many filters

-find out the Tenessee is not the bottom 10 states for nutrition intake. However, it still wont meet the guideline for 3 cups of vegetable/fruit intake per day.

<https://www.cdc.gov/media/releases/2017/p1116-fruit-vegetable-consumption.html#:~:text=Depending%20on%20their%20age%20and,of%20a%20healthy%20eating%20pattern>.

-TN is the top 10 states of obese

-Question col filter ‘Overweight or obesity among adults aged >= 18 years’ & ‘Overweight or obesity among high school students’, ‘Overweight or obesity among women aged 18-44 years’

- Overweight or obesity among high school students and ‘Overweight or obesity among women aged 18-44 years’ are also top 10 and is in rising trend

-Overall, will string containing meeting physical activities, TN is bottom 10. so, it is consistent that is the top 10 obese states.

11/30

I found out that I wrongly saved excel file into a csv file format and the copy I created from the master is gone. No charts are shown from the exploratory analysis.

12/1/20

Troubleshooted that I need to save the excel file into excel workbook so I can manipulate the data and keep exploring in excel. At this point, it is easier to explore the data in excel as there are many filters in the data.

I can see that for understanding the relationship % obese and vegetable intake / fruit intake less than one daily. The survey for vegetable intake/fruit intakes only have 2017 so I use overall in stratficiation column in file Nutrition\_\_Physical\_Activity\_\_and\_Obesity\_\_Behavioral\_Risk\_Factor\_Surveillance\_System.copy.

I will look at physical activity and possibly other stratification to see if there is a difference.

12/2/20

Problem is excel is slow to load data.

I have answered question 1 in file U.S.\_Chronic\_Disease\_Indicators\_\_CDI\_.copy

12/12/20

I am analyzing and plotting the data values points on tableau map (using mapbox style map in the background). I have background information of the 500 cities which are the 500 largest US cities and approximately 28,000 census tracts within these cities.

<https://www.cdc.gov/places/methodology/index.html>

Methodology: The multilevel regression and poststratification (MRP) approach is flexible and will help CDC provide modeled estimates of the prevalence for each indicator at the census tract and city levels.

<https://www.cdc.gov/places/measure-definitions/unhealthy-behaviors/index.html#obesity> 🡪 good to introduce in the beginning what your data sources are

Racial and Ethnic Approaches to Community Health

<https://www.cdc.gov/nccdphp/dnpao/state-local-programs/reach/index.htm>

<https://www.cdc.gov/nccdphp/dnpao/state-local-programs/reach/pdf/REACH-overview-2017-508.pdf> one example of reach program including how it works and what is the impact.

12/17/20

Prepare presentation flow

**Outline**

**-Executive Summary & Motivation**

**-Data**

-Five datasets from Centers For Dicease Control and Prevention(CDC)

-Look at national, state and city levels CDIs between 2011 -2018

**-Scope of the Project**

-Definition of Obesity:

Body Mass Index (BMI) is a person’s weight in kilograms divided by the square of height in meters. A high BMI can be an indicator of high body fatness.

BMI: 30-99.8

-Life Style Factors: Physical Activity, Fruits and Vegetables intakes

-Measurement Unit: Crude Prevalence

-Survey: Behavioral Risk Factor Surveillance System (BRFSS)

**-Assess Health Issues**

-Overall Health in CDIs between states

-Obesity Ranking between states

-Social Descriptors Breakdown

**-Prioritize needs**

-Lifestyle Factors

-No activity

-Comparision between different lifestyle factors

-Coorelation between different lifstyle factors

-Neighborhood maps of largest cities in TN

-matched obesity and no activity population

**-Take Action**

-shape policies, programs, and neighborhood activities that address health needs and promote a high quality of life for all

**Export html file from python into a URL using window 10 (training by Mark)**

Search box:

-type ‘ turn on windows’, open and IFS (check box)

-open file explorer, inetpub, wwwroot is the local web location and you can copy your site path in this root folder

-open web browser then type in ‘localhost/file path.html and you see the webpage and it can run as http

-go to search box type iis manager, default eg. Google chrome, under action you can click restart when your website is not working

12/24/20

-Formatting and double-checking data type in data source, and social descriptor graph issues as well as folium map city names conflict

- Decided to change lifestyle factors to leading health indicators under scope of projects (not social determinants of health as I believe social determinants need to be investigated as to why certain people have lower physical activity level i.e. availability of gym facilities?)

Reference: <https://www.healthypeople.gov/2020/leading-health-indicators/2020-lhi-topics/Nutrition-Physical-Activity-and-Obesity>

1/4/20

Definition supporting what crude prevalence is

It is the health problem divided by total population and you compare to other group like state per 10000 population. It is a general view on which group is riskier.

<https://sphweb.bumc.bu.edu/otlt/MPH-Modules/EP/EP713_StandardizedRates/EP713_StandardizedRates2.html>

I am not an epidemiologist and further analysis might need to tackle for specific age rate adjusted rate to look at further difference in rate in comparison.

<https://www.cdc.gov/cancer/uscs/technical_notes/stat_methods/rates.htm> supporting crude rate importance in understanding the burden of specific group in a population

for the social determinants of health, examining the data source, i.e age group and crude prevalence filters applied, each age group has a sample size in a specific year and how many people are obese in that year divided by the at risk population times the 10000.

1. Information provided by <https://www.vumc.org/pophealth/> [↑](#footnote-ref-1)
2. Information provided by <https://www.healthypeople.gov/2020/leading-health-indicators/2020-lhi-topics/Nutrition-Physical-Activity-and-Obesity> [↑](#footnote-ref-2)