Ruth Maina

DSC 680 - Applied Data Science

Professor Amirfarrokh Iranitalab

October 12, 2024

A Comprehensive Exploration of the Obesity Epidemic

Final White Paper

Topic: A Comprehensive Exploration of the Obesity Epidemic

Background/History:

The Center of Disease Control (CDC) estimates about forty percent of the American population is currently obese. As the years go by, some diseases which have been on the rise have been attributed to individuals being overweight.

Business Problem:

Back in the 70s, the obesity rate in America used to be in the teens. In a span of about 50 years, the rate has continued to steadily increase, and it is projected to reach 50 percent by the year 2030. The purpose of this project is to explore obesity data for the last ten years, to let the data tell a story on how this is trending and to perform a prediction of where this could go in the next few years. I'm also curious to see if the growth prediction value of 50 percent by 2030 is accurate.

Data Explanation:

Dataset:

- I chose a dataset from the Center for Disease Control and Prevention (CDC) – it has 602 observations and 43 variables, which is small but a good size to work with. Below are the column details of the dataset

#	Attribute	Description
1	YearStart	Year survey taken
2	LocationAbbr	State short name
3	Data_Value	Obesity Percentage
4	Sample_Size	Number of Respondents
5	Age (years)	Age in Years
6	Education	Education level
7	Gender	Gender level
8	Income	Income level
9	Race / Ethnicity	Race or Ethnicity

Data preparation steps:

- For this study, I chose to focus on adults' obesity data, even though there's data available for other age groups such as adolescents, two- four-year-olds, as well as 3–23-month-olds.
- Of all the available variables, I selected to perform key calculations on the obesity percentage field, which comes in 3 sub-fields containing obesity percentages - an actual value, a high confidence limit and a low confidence limit. I performed calculations only on the actual value.
- I loaded the csv data into an R data frame and performed cleanup activities such as checking for null values, unwanted characters, and for missing values to enforce data integrity. All the aforementioned activities yield clean data, which impacts model performance positively:
    - A few field values in the target variable contained a character value of '~' which is an irregularity so these we're omitted.
    - Also omitted NA values
    - Checked the variable data type and noticed it was stored as 'chr' - converted this to integer values so calculations could yield correct results.

Methods:

- With the 3 main variables, namely Year, State, and Obesity percentage/Data Value, I fit a linear model using the Year variable as the predictor and Obesity as the outcome. A linear model is best suited for the prediction of a target variable based on one or more input features, which is my exact use case.
- I also explored the other variables such as State values, Yearly values, and so on for a better understanding of the obesity impact - *Refer to the Appendix section for more information.*

Analysis:

- To get a feeling of the data distribution, I used a density plot, a histogram, and scatter plots to visualize the data.
- The distribution of obesity was mostly normal/symmetrical and unimodal with one clear peak in the data, without outliers or a skew.

- I also used additional scatterplots view Obesity by State which revealed three states with the highest percentage being Kentucky, Mississippi, and West Virginia respectively, all around 40% value - *Refer to the Appendix section for the plots/visualizations*
- I predicted obesity for the next 10 years and the outcome indicated a steady overall increase but not more than 42%. This number was surprising as I anticipated a high increase, but it appears prediction indicates a leveling off which is good news compared to a continual upward increase.

Conclusion:

- The above analysis has shed light on how obesity is trending, providing analysis of the past decade and a potential prediction of the next several years. While the last decade/actual data shows a sharp increase, the latter decade prediction shows a levelling off, which could mean that some causative factor could be helping avoid gradual increases compared to the last decade.
- Overall, the issue can be tackled in many ways such as exercise, and intervention opportunities could be explored from all angles, particularly from governing bodies and from community education.

Assumptions:

- Underlying obesity causes are not explored as part of this study.
- It is assumed that the CDC data is accurate and reliable for the study.

Limitations:

- The model accuracy could be increased by adding additional years of data. For example, instead of starting at 2011. We could go back several more years to fine tune the prediction. Bigger samples could potentially enable smaller margins of error.

Challenges:

- This is an independent prediction, and testing is not possible for this scenario. Furthermore, the results of such a test would also not be foolproof as every human body is different thus would react differently/yield slightly different results. These details will be communicated to the users of the model.

Future Uses/Additional Applications:

- The model may be used for academic research, it can be used by clinicians to inform individuals on the subject, and overall, by health-conscious individuals for awareness and caution. The trend percentages for the last decade and future prediction indicate an increased disparity from 2011 – 2032, which would help the audience see the importance of the matter.

Recommendations:

- One recommendation is to resample and rerun the model, as well as using data from a different age group to observe the trends, for comparison and general informational purposes. The analysis has proven that obesity is a weighty subject that should be tackled collectively by all.
- A deeper understanding on the issue can be gained by analyzing potential causes, especially on the highest-ranking states.

Implementation Plan:

- The model will first be deployed into lower testing/sandbox environments for validation before it goes live into production. The Site Reliability teams, or production support will be trained on the model details and will be given a troubleshooting runbook to enable model support is covered post deployment. This will ensure that any breaks or unexpected behavior is addressed promptly.

Ethical Assessment:

- It is imperative that the predicted outcome is accurate since putting incorrect recommendation or providing less than precise correlation could pose ethical implications. The core data is from CDC, a very reputable source ideal for health information. End users will be made aware of the entire process for transparency. Privacy laws is another consideration as the data pertains to health information. Disclosing personal information is not permitted.

Questions and Answers:

1. How was the data gathered?
   The data was collected from individuals who opted-in to share part of their results during annual wellness check-up with their doctors, as part of a clinical study.
2. If this is real human data, did the data owners approve of this use?
   Yes the individuals approved by opting-in and full disclosure occurred regarding the intent and how the data was going to be used
3. How can you ensure privacy is adhered to?
   Our networks are fully firewall secured, and two-way authentication is enabled for access, which is also restricted to individuals on a need-to-know basis.
4. What data retention policy will be applied?
   Our intent is to keep the data for five years, after which it will go through destruction process (complete purge)
5. What policies are consulted to ensure compliance with health regulations?
   - *HIPAA (Health Insurance Portability and Accountability Act) was consulted to ensure patient data safety*
   - *Centers for Disease Control and Prevention (CDC) was consulted for additional reference*
6. How was the sampling done?

The 80/20 Pareto Principle was used to split the data into training and test datasets accordingly – 80% used for training the model mainly to learn the correlation and the relationship between the data, and 20% used for model validation.

7. Was any of the gathered data not used?

   The entire dataset was used, except for the identifier column

8. How do you plan to use the model?

   The model will be made available to healthy conscious people interested in the subject, and for academic research

9. Why was the model chosen over others?

   The linear model is best suited for prediction of a target variable based on one or more input features, which is my exact use case

10. Can the model be applied to all age groups?

    Yes this model may be applied to all age groups

GitHub Code link: https://github.com/ruthmaina2022/Data-Science-Portfolio/blob/main/Applied%20Data%20Science%20Projects/Project%202/Ruth%20Maina_DSC680_Assignment8.2_Project%202_Milestone%203_Final%20Report%20Code.ipynb

GitHub Project2 link: https://github.com/ruthmaina2022/Data-Science-Portfolio/tree/main/Applied%20Data%20Science%20Projects/Project%202

References:

Behavioral Risk Factor Surveillance System (2020, November 10). Data Catalog.

https://data.cdc.gov/api/views/hn4x-zwk7/rows.csv?accessType=DOWNLOAD

CDC (n.d.) Nutrition, Physical Activity, and Obesity: Data, Trends and Maps.

https://nccd.cdc.gov/dnpao_dtm/rdPage.aspx?rdReport=DNPAO_DTM.ExploreByTopic&islClass=OWS&islTopic=&go=GO

National Center for Health Statistics (n.d). Health, United States.

https://www.cdc.gov/nchs/hus/data-finder.htm.

Global Health Data Exchange (n.d). United States Physical Activity and Obesity Prevalence by County 2001-2011.
https://ghdx.healthdata.org/sites/default/files/record-attached-files/IHME_USA_OBESITY_PHYSICAL_ACTIVITY_2001_2011.csv

CDC (n.d.). Causes of Obesity

https://www.cdc.gov/obesity/basics/causes.html

CDC (n.d.). Adult Obesity Facts

https://www.cdc.gov/obesity/data/adult.html

World Health Organization. (2021, June 9). Obesity and overweight.

https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight

National Geographic. (2017, April 6). 5 “Blue Zones” Where the World’s Healthiest People Live

https://www.nationalgeographic.com/books/article/5-blue-zones-where-the-worlds-healthiest-people-live

Dr.Mandal, A MD. (2019, February 27). Obesity and Fast Food.

https://www.news-medical.net/health/Obesity-and-Fast-Food.aspx

NCHS Data Brief No. 322 (2018, October). Fast Food Consumption Among Adults in the United States, 2013–2016.

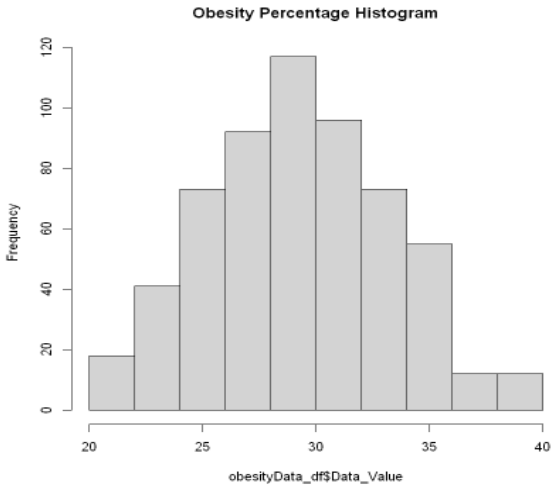
https://www.cdc.gov/nchs/products/databriefs/db322.htm#:~:text=In%202013%E2%80%932016%2C%2036.6%25,adults%20aged%2060%20and%20over.

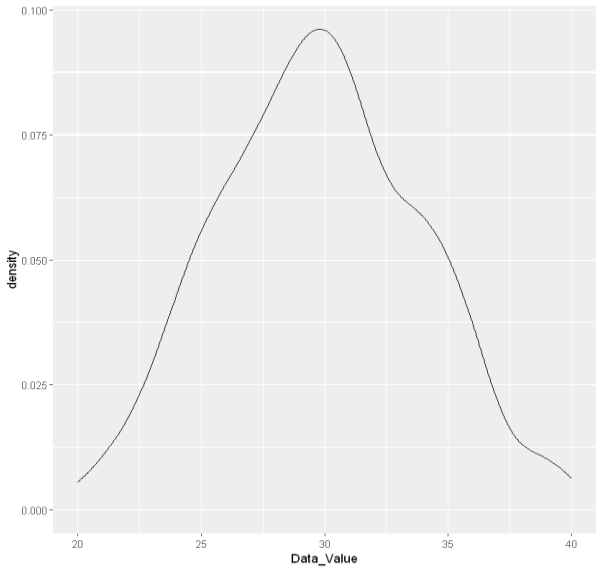
Wikipedia (n.d.). Obesity in the United States.

https://en.wikipedia.org/wiki/Obesity_in_the_United_States#Contributing_factors

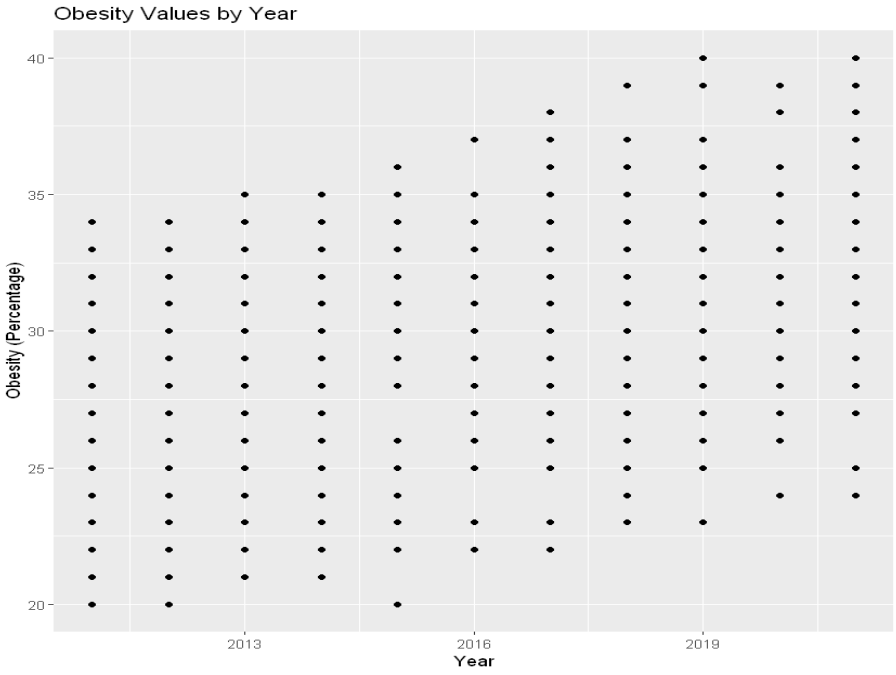
Appendix:

- Per below density plot and histogram, the distribution of obesity data is mostly normal/symmetrical and unimodal with one clear peak in the data, without outliers or skew. Most of the population centers around the mean – this distribution could accurately be used to model obesity.





- A scatterplot view by Year below reveals a steady increase as expected peaking at 40% value.



- A scatterplot view by State reveals three states with the highest percentage being Kentucky, Mississippi, and West Virginia respectively, all around 40% value.