OBESITY EPIDEMIC IN THE USA

FIT5147 Data exploration project report

MD SAADMAN HOSSAIN
STUDENT ID: 31043313
TUTOR: MOHAMMAD HAQQANI

INTRODUCTION

This project aims to analyse and visualise the obesity epidemic in USA to find trends and relations. Particular points of interest include states wise analysis of obesity data, causes of obesity in the US population and insights towards reducing this ever-growing socio-economic problem by analysing the behavioural risk factors associated with obesity.

I would like to address the following questions through exploration of the dataset of choice:

- Which states in the USA have the most and least obese/overweight population?
- Do activity levels if individuals affect the level of obesity in the US population?
- Does age or race have an effect on the obesity levels?
- Are there any differences in levels of obesity by gender? Why is this the case?
- Which factors are responsible for an unhealthy diet? Is there any relation between unhealthy food habits and obesity?
- Do social and economic conditions affect obesity levels?

Source of the data set-

Nutrition, physical activity and obesity – Behaviour Risk Factors Surveillance System. The data is collected by Centers for disease Control and Prevention and can be found in the USA government federal database (public use).

DATA WRANGLING

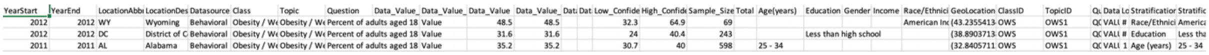
CDC Data:

1. Firstly, I downloaded the data set in .csv file format from the Data.gov website which is the federal website for databases relating to the US government. This data set contains raw data about the behavioural risk factors associated with nutrition, physical activity and obesity. The raw data set consists of 33 unique attributes (columns) and around 60,000 rows (contains data for year from 2011 to 2018). The attributes include information such as starting year, end year, location description, class of the questions, questions, data value(percentage of each question), sample size, socio-economic factors (age, gender, education, income, race), geolocation(latitude and longitude), stratificationcategory (describing the socio-economic factors).

2. I have used data for all the years and for all the states of USA including the territories that are inhabited. I then proceeded to load the raw data into R for wrangling.

3. Libraries used along with built-in R functions were tidyverse, tidyr, compare, dplyr which were a great help in the wrangling process

4. The raw data set is quite messy with poor attribute names (shown in the picture below). Appropriate filtration was conducted to reformat the data.



YearStart	YearEnd	LocationAbbı	LocationDes	Datasource	Class	Topic	Question	Data_Value_	Data_Value_	Data_Value	Data_Value_	Dat:	Dat:	Low_Confide	High_Confide	Sample_Size	Total	Age(years)	Education	Gender	Income	Race/Ethnici	GeoLocation	ClassID	TopicID	Qu	Data	Lo	Stratificatior	Stratific
2012	2012	WY	Wyoming	Behavioral	Obesity / We	Obesity / We	Percent of adults aged 18		Value	48.5	48.5			32.3	64.9	69						American Inc	(43.2355413	OWS	OWS1	QC	VALU	#	Race/Ethnici	America
2012	2012	DC	District of C	Behavioral	Obesity / We	Obesity / We	Percent of adults aged 18		Value	31.6	31.6			24	40.4	243			Less than high school				(38.8903713	OWS	OWS1	QC	VALU	#	Education	Less tha
2011	2011	AL	Alabama	Behavioral	Obesity / We	Obesity / We	Percent of adults aged 18		Value	35.2	35.2			30.7	40	598		25 - 34					(32.8405711	OWS	OWS1	QC	VALU	1	Age (years)	25 - 34

5. Each row in this data set contain yearly data for a particular state relating to the question attribute (class attribute has 3 levels and each levels of the class attribute is associated with more levels in the question attribute i.e. percent of adult with obesity, percent of adult with no physical activity etc). the Stratificationcategory1 attribute determines the socio-economic factors which is associated with the data_value.

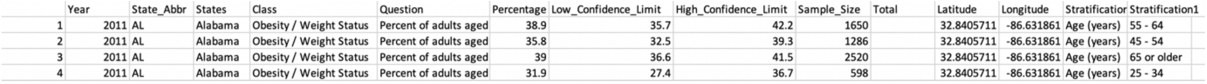
6. After reading the data into R, I ordered it by the location and year.

7. Next step was to remove the parenthesis and comma in the geolocation column and separated it into 2 columns (latitude and longitude) for the purpose of plotting a map for my visualisation.

8. Then I proceeded to remove the unwanted columns (this includes ID columns and also socio-economic factor columns as they are referrable through the stratification column. I also use R to check if the start and end year were identical columns, as they were identical, I compressed it into one column.

9. Around 10 percent of the percentages in the Data_value column were null; I will explain how I deal with the missing values in the data checking section of the report.

10. With the remaining columns I have used the dplyr library to rename these columns accordingly. Then I wrote the cleaned and formatted data set into a csv file for the use of exploration and visualisation. The cleaned data set now looks like this:



	Year	State_Abbr	States	Class	Question	Percentage	Low_Confidence_Limit	High_Confidence_Limit	Sample_Size	Total	Latitude	Longitude	Stratificatior	Stratification1
1	2011	AL	Alabama	Obesity / Weight Status	Percent of adults aged	38.9	35.7	42.2	1650		32.8405711	-86.631861	Age (years)	55 - 64
2	2011	AL	Alabama	Obesity / Weight Status	Percent of adults aged	35.8	32.5	39.3	1286		32.8405711	-86.631861	Age (years)	45 - 54
3	2011	AL	Alabama	Obesity / Weight Status	Percent of adults aged	39	36.6	41.5	2520		32.8405711	-86.631861	Age (years)	65 or older
4	2011	AL	Alabama	Obesity / Weight Status	Percent of adults aged	31.9	27.4	36.7	598		32.8405711	-86.631861	Age (years)	25 - 34

Data Checking:

Missing Data:

In the CDC dataset, about 10 percent of the rows contain missing values in the Data_value column. There also was a lot of missing values in some other attributes which was dealt with by deleting the columns (data not important for analysis). I used tableau to visualise the missing values aggregated by US states (abbreviations used) and the result is shown below:

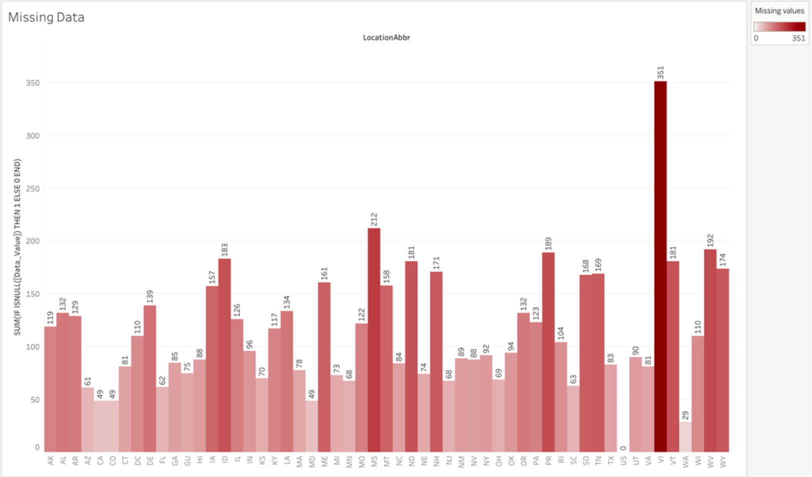


Figure 1: bar plot for missing values

As we can see there isn't any discernible pattern in the missing values. Only observation I was able to come up with is that there isn't any missing data for US as a whole, and Virgin Islands had the most amount of missing values. There is a column in the raw dataset (Data_Value_Footnote) in which it states that the missing values are due to low sample size. This confirms my observation that the missing data was missing at random (MAR). I chose to delete the rows with the missing values as the data is pretty much missing at random (deleting the rows will barely affect the percentages).

Outliers in data:

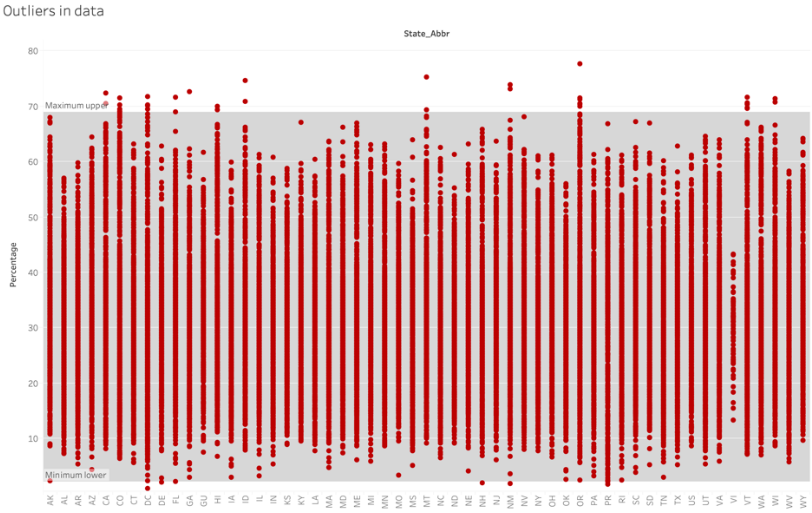


Figure 2: Dot plots represented by confidence intervals

To check for outliers in the data, I plotted all the data points by state and calculated 95 percent confidence intervals (percentage +/- 2*standard deviation) to check for outliers. As seen in the figure above, most of the data falls inside the confidence interval. Although some of the data is considered as outliers, I chose to keep this data, as outliers maybe be caused by various factors(lower sample size i.e. if we look at both figures we can see that virgin islands have missing values due to smaller sample size but the data doesn't vary as much as the percentages have low variance) and not necessarily due to errors in data. In this data the observed data is about the states in USA, where cultures, values and lifestyles are different for each state. Thus, the outliers in the data can be considered as a normal occurrence.

Data Exploration:

Most and least obese/overweight representation according to US states

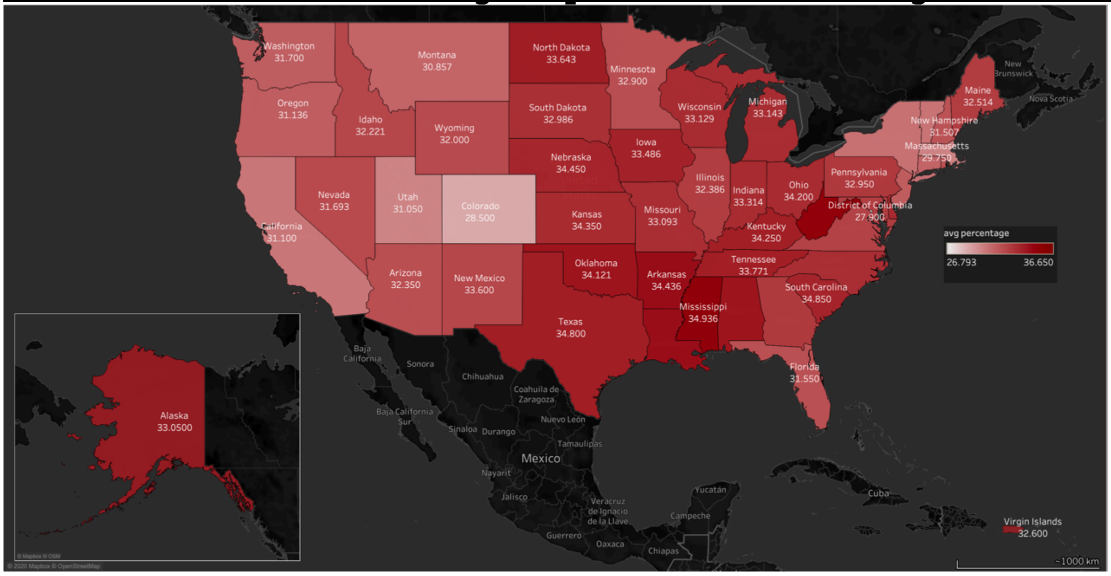


Figure 3: choropleth map for obesity averages by state

My first point of interest was to find out the obesity levels in the USA represented by state. To find answers, I had to further wrangle the formatted data set. I used R to order the data percentages according to the class "obesity/weight status". Then I proceeded to aggregate the data percentages by state names and with the appropriate questions (a column in my data which represents levels) relating to obesity. Then I used tableau to generate a choropleth map to show the percentages as hue of red colour (to indicate it's a matter of concern). I used the geolocation data to plot the us map.

From the figure above we can see that West Virginia, Mississippi and Arkansas are the most obese states in the US, whereas Colorado is least obese state in US.

Do activity levels affect obesity levels in the US population?

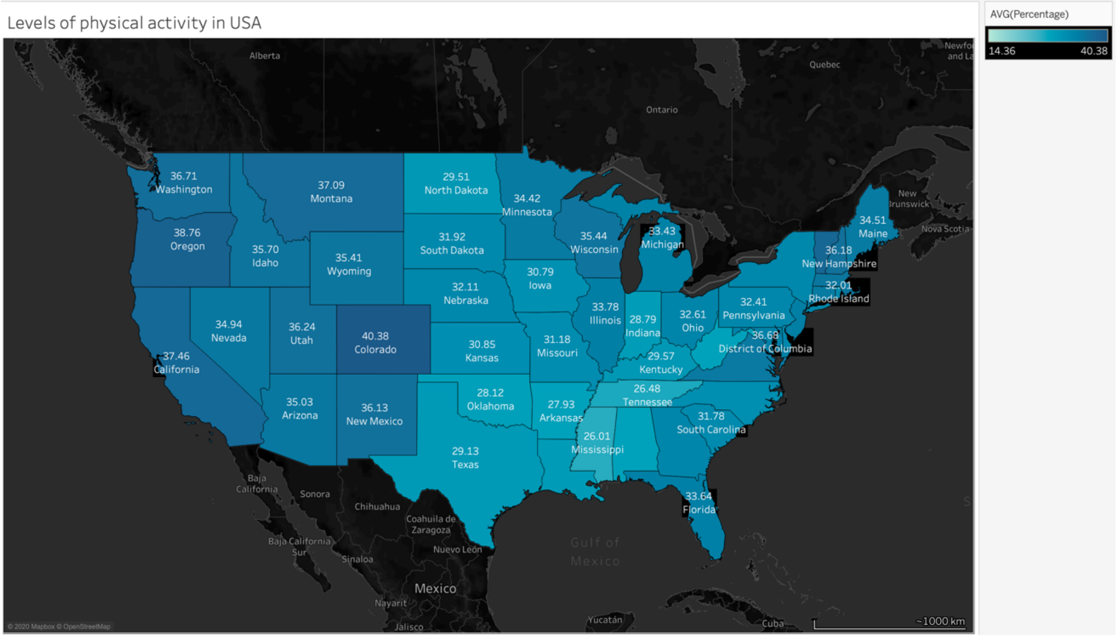


Figure 4: Choropleth map for physical activity levels

For this question, I did a similar process of wrangling with R as question 1(ordering and aggregation). From the figures 3 and 4 above we can see that Colorado has the highest level of physical activity. If we compare this information to the figure from question one, it's also quite interesting to note that Colorado has the lowest level of obesity amongst the US states. Mississippi and Arkansas have low levels of physical activity and consequentially are on the high spectrum in obesity levels. We can see this trend for each US state (I only highlighted the extreme cases). Through analysis of this trend I can safely say that physical activity and obesity level are inversely correlated (states with higher levels of physical activity have low levels of obesity).

Effect of race and age on obesity

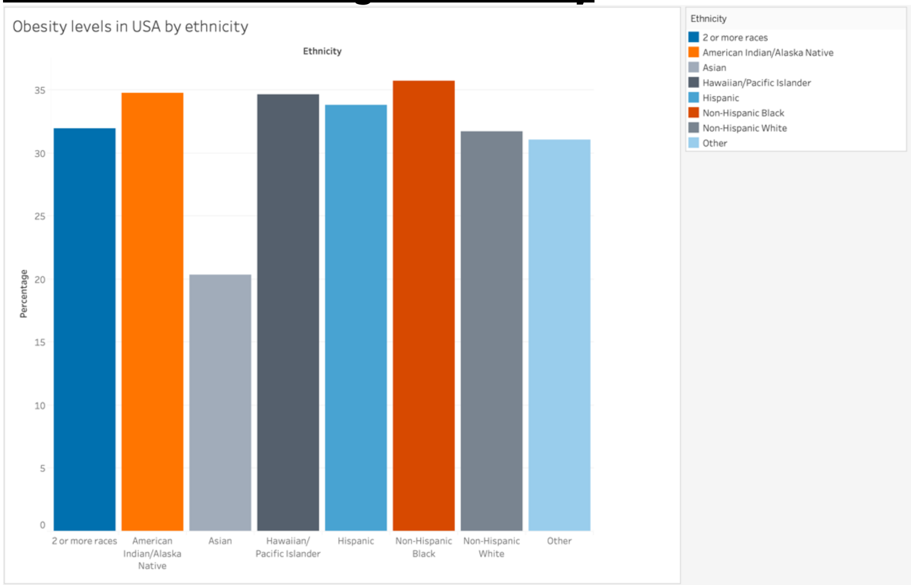


Figure 5: Bar plot for Obesity by ethnicity

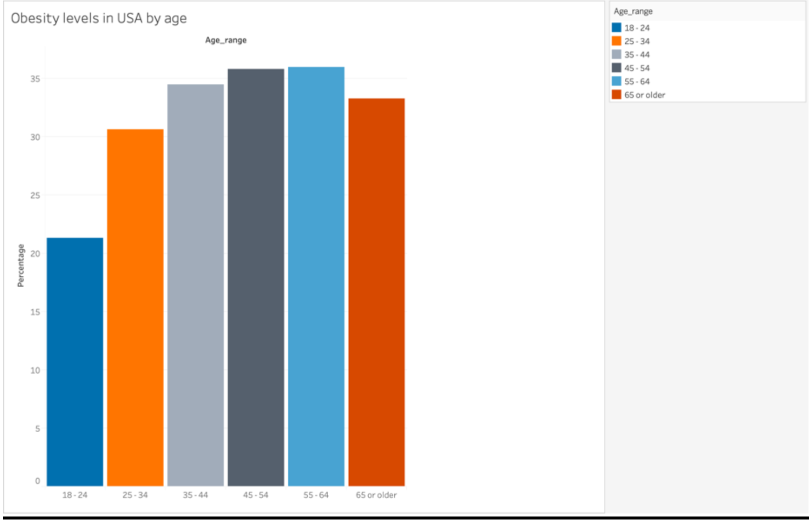


Figure 6: Bar plot for obesity by age

We can see from figure 5 above that obesity levels amongst different races vary. Non-Hispanic blacks and native Americans are amongst the most obese races and Asians have the lowest level of obesity.

Obesity levels also vary according to age levels. From figure 6 we can see that young people in the age bracket of 18-24 have the least obesity. Obesity levels gradually increase as people age and it reaches its peak when people are around 55-64 years old. One interesting thing to note is that obesity levels drop when people get older than 65 years of age. This may happen due to higher mortality rate in that age group.

Are there any differences in levels of obesity by gender? Why is this the case?

Firstly, I ordered the wrangled data by year and stratification in R. Then I aggregated the data frame based on gender. Ggplot2 library was used to visualise this data frame into the figure 7 below. The figure is a grouped bar chart and aims to show the levels of obesity by gender and year. We can see that obesity levels are higher amongst males than females. A particular point of interest arises when we look at the rate at which percentage of obese population increases for the genders. Females are becoming obese at a higher rate than males. But why is this the case?

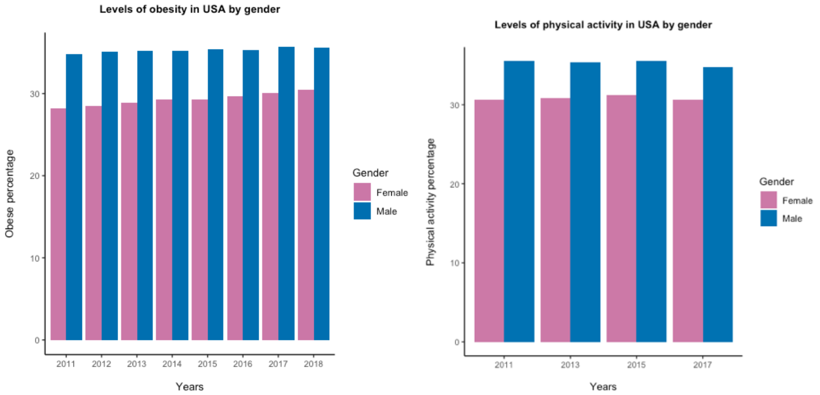


Figure 7: ggplot2 grouped bar plots - Obesity by gender and physical activity by gender

We can see from the figure that physical activity levels are also different amongst males and females. Males are taking part in muscular training and leisure time physical activity more than females. This might be a factor that helps to answer the question. My hypothesis: obesity rate amongst females is increasing faster than males due to the fact that males are on average more engaged in physical activity than females.

Factors influencing an unhealthy diet

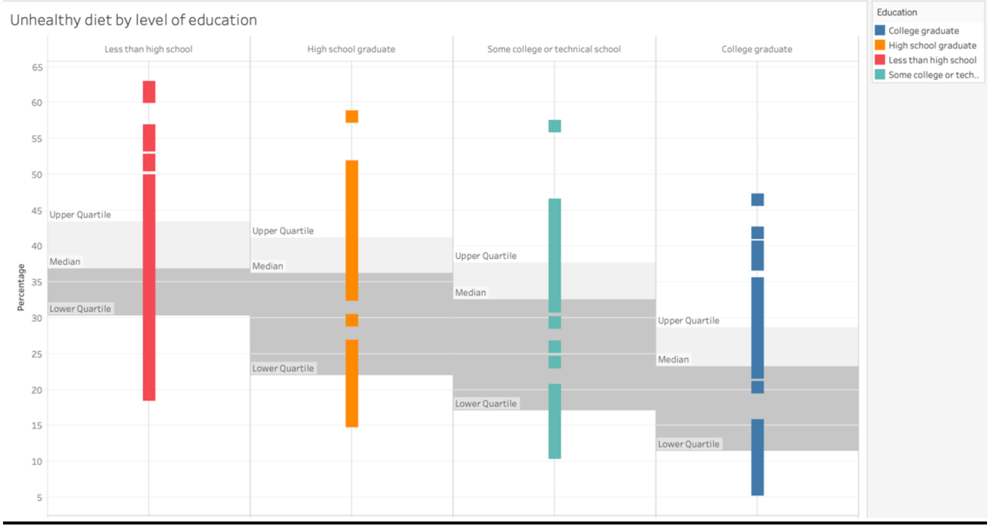


Figure 8: Boxplot for unhealthy diet by education

Figure 8 above shows boxplots of unhealthy diet by level of education. I defined unhealthy diet based on the data available which includes percentages of people who consume fruits and vegetables less than 1 time daily. We can see that unhealthy diet is related to level of education. In the US population, people who have an education level of less than high school are more prone to following unhealthy food habits whereas people with a college degree and less likely to have an unhealthy diet. From this analysis, my conclusion is that humans adapt to a healthier lifestyle as they become more educated.

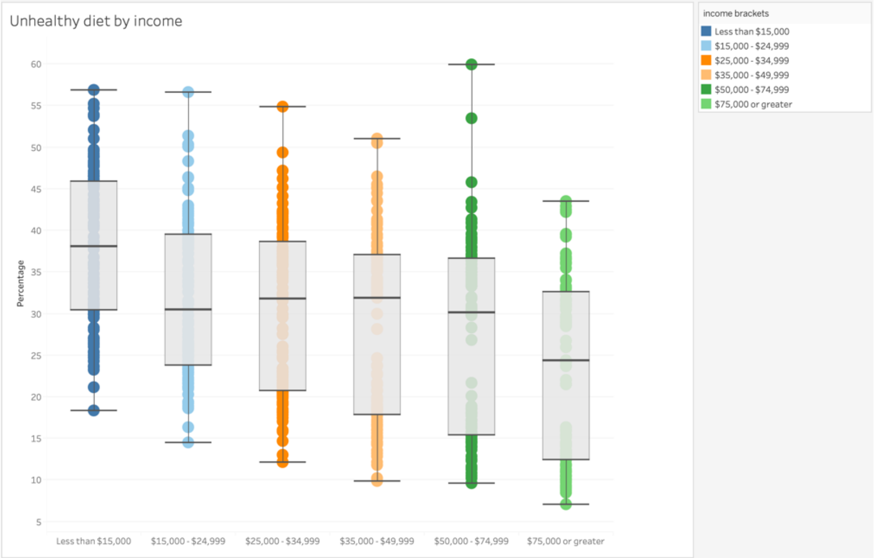


Figure 9: box and whiskers for unhealthy diet by income

From figure 9 above (box and whiskers plot) we can see that unhealthy diet percentage gradually drops as people start earning more. At the lowest income bracket of less than $15000, highest level of unhealthy food habits and at the highest income bracket of $75000 and greater we can see people are more health conscious (lowest mean and variance). The middle brackets have similar means although the percentages vary quite a bit. The huge variance in the middle classes is observed due to most of the population falling in this category.

From the figures above it is safe to conclude that income and education are quite reliable indicators of unhealthy food habits.

Relation of unhealthy diet and obesity:

Figure 10 shown below was created using ggplot2 in R. It's a scatterplot of obesity percentages in relation to unhealthy food habits. I fitted a regression line (loess method in ggplot2) through the points to highlight the positive correlation between obesity and unhealthy diet. We can clearly see that obesity levels rise as people adapt unhealthy food habits.

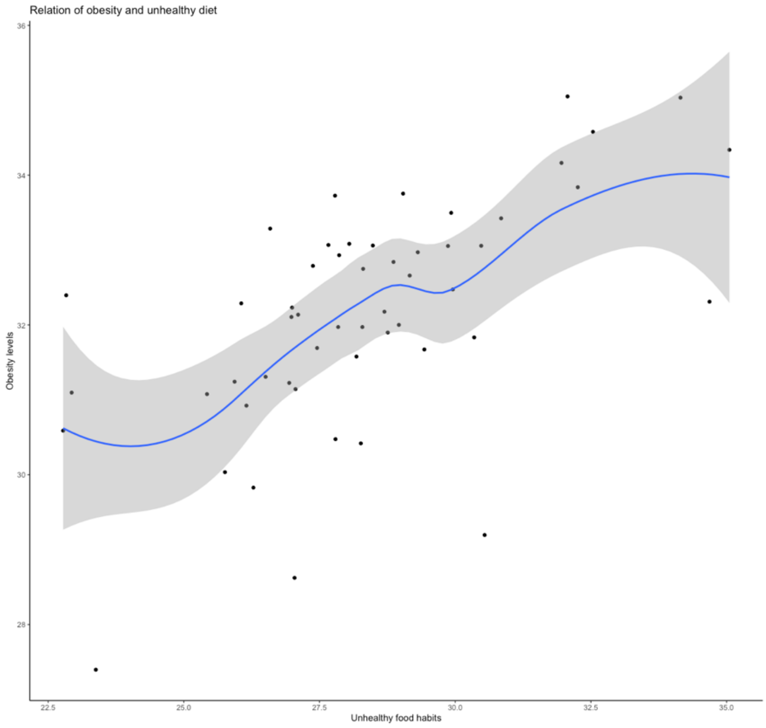
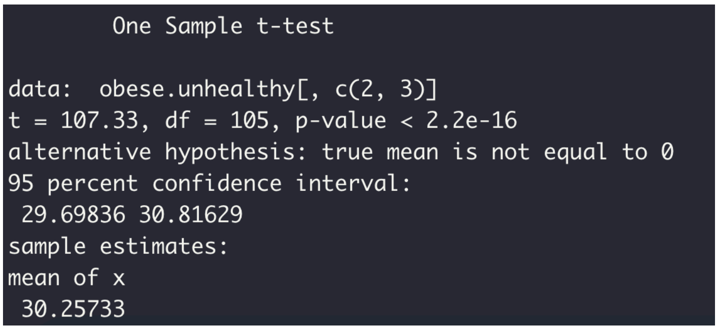


Figure 10: ggplot2 scatter plot for obesity vs unhealthy diet

T-test for obesity vs unhealthy diet



```
        One Sample t-test

data:  obese.unhealthy[, c(2, 3)]
t = 107.33, df = 105, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 29.69836 30.81629
sample estimates:
mean of x
 30.25733
```

I performed a statistical analysis(t-test) for obesity vs unhealthy diet using R. results indicate a p-value which is less than 0.05. this indicate that the effect of unhealthy diet on obesity is indeed significant and my visualisation is also valid.

Do social and economic conditions affect obesity levels?

Socio-economic conditions include age, gender, race, education and income. From the figures above we saw that race, age and gender are influential factors in obesity levels. Although income and education doesn't seem to affect obesity levels as much. Income and education are not reliable indicators of obesity from my analysis.

In the case of unhealthy food habits, it might not be an individual's choice that dictates if they are able acquire healthy foods for consumption (low income and insufficient education may lead to poor decisions or consumption of low-quality foods). On the other hand, obesity seem to be a rather preventable disease in most cases (small percentage of people are obese due to pre-existing health conditions.) which means it is mostly dictated by an individual's choices. Thus, their level of education and income has little effect on decisions that lead to obesity.

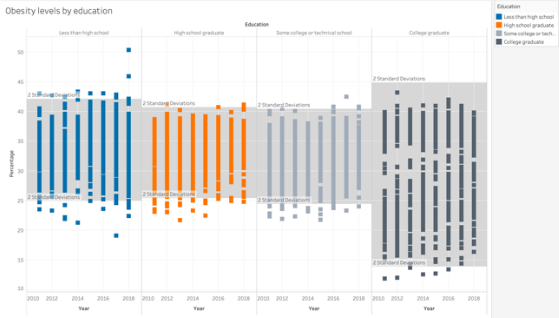


Figure 11: obesity by education

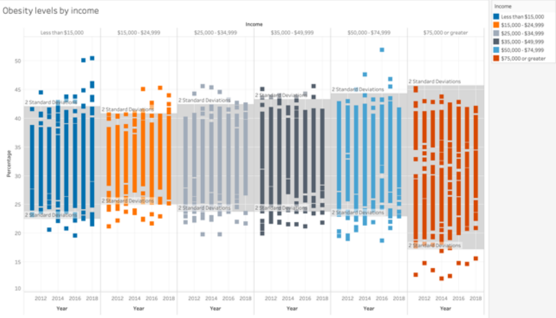


Figure 12: obesity by income

Conclusion:

By analysing the obesity data for USA, I was able to find trends and insights from the data such as females are increasingly becoming obese compared to males, West Virginia and Mississippi are the most obese states in USA, Colorado as state is more into physical activity, income and education are good indicators of unhealthy diet. I also learned that obesity and unhealthy food habits are correlated, race and age are good indicators of obesity whereas income and education is not. In my opinion I was able to find some evidence to support the questions with my visualisations, although there is always room for improvement.

Reflection:

I have learned a lot about how obesity is affecting the US population and is turning into a global pandemic during this project. Although some socio-economic factors are key indicators, ultimately it is a preventable disease. I tried to do my best to extract information from the data set, but I would have liked to link the obesity issue with national health spending of USA. I think exploring how the obesity pandemic is affecting the US national health care is an interesting topic.

Bibliography:

Dataset used:
https://catalog.data.gov/dataset/nutrition-physical-activity-and-obesity-behavioral-risk-factor-surveillance-system