UN Nutrition Analysis

Sathvik Maridasana Nagaraj

2024-07-21

# Import the CSV file  
Nutrition\_Physical\_Activity\_and\_Obesity <- read.csv("D:/BioInformatics/US Nutrition Analysis/Nutrition\_\_Physical\_Activity\_\_and\_Obesity.csv")  
  
# View the first few rows of the dataset  
head(Nutrition\_Physical\_Activity\_and\_Obesity)

## YearStart YearEnd LocationAbbr LocationDesc  
## 1 2020 2020 US National  
## 2 2014 2014 GU Guam  
## 3 2013 2013 US National  
## 4 2013 2013 US National  
## 5 2015 2015 US National  
## 6 2015 2015 GU Guam  
## Datasource Class  
## 1 Behavioral Risk Factor Surveillance System Physical Activity  
## 2 Behavioral Risk Factor Surveillance System Obesity / Weight Status  
## 3 Behavioral Risk Factor Surveillance System Obesity / Weight Status  
## 4 Behavioral Risk Factor Surveillance System Obesity / Weight Status  
## 5 Behavioral Risk Factor Surveillance System Physical Activity  
## 6 Behavioral Risk Factor Surveillance System Physical Activity  
## Topic  
## 1 Physical Activity - Behavior  
## 2 Obesity / Weight Status  
## 3 Obesity / Weight Status  
## 4 Obesity / Weight Status  
## 5 Physical Activity - Behavior  
## 6 Physical Activity - Behavior  
## Question  
## 1 Percent of adults who engage in no leisure-time physical activity  
## 2 Percent of adults aged 18 years and older who have obesity  
## 3 Percent of adults aged 18 years and older who have obesity  
## 4 Percent of adults aged 18 years and older who have an overweight classification  
## 5 Percent of adults who achieve at least 300 minutes a week of moderate-intensity aerobic physical activity or 150 minutes a week of vigorous-intensity aerobic activity (or an equivalent combination)  
## 6 Percent of adults who achieve at least 150 minutes a week of moderate-intensity aerobic physical activity or 75 minutes a week of vigorous-intensity aerobic physical activity and engage in muscle-strengthening activities on 2 or more days a week  
## Data\_Value\_Unit Data\_Value\_Type Data\_Value Data\_Value\_Alt  
## 1 NA Value 30.6 30.6  
## 2 NA Value 29.3 29.3  
## 3 NA Value 28.8 28.8  
## 4 NA Value 32.7 32.7  
## 5 NA Value 26.6 26.6  
## 6 NA Value 27.4 27.4  
## Data\_Value\_Footnote\_Symbol Data\_Value\_Footnote Low\_Confidence\_Limit  
## 1 29.4  
## 2 25.7  
## 3 28.1  
## 4 31.9  
## 5 25.6  
## 6 18.6  
## High\_Confidence\_Limit Sample\_Size Total Age.years. Education  
## 1 31.8 31255   
## 2 33.3 842 High school graduate  
## 3 29.5 62562   
## 4 33.5 60069   
## 5 27.6 30904   
## 6 38.5 125   
## Gender Income Race.Ethnicity GeoLocation ClassID  
## 1 Hispanic PA  
## 2 (13.444304, 144.793731) OWS  
## 3 $50,000 - $74,999 OWS  
## 4 Data not reported OWS  
## 5 Less than $15,000 PA  
## 6 Hispanic (13.444304, 144.793731) PA  
## TopicID QuestionID DataValueTypeID LocationID StratificationCategory1  
## 1 PA1 Q047 VALUE 59 Race/Ethnicity  
## 2 OWS1 Q036 VALUE 66 Education  
## 3 OWS1 Q036 VALUE 59 Income  
## 4 OWS1 Q037 VALUE 59 Income  
## 5 PA1 Q045 VALUE 59 Income  
## 6 PA1 Q044 VALUE 66 Race/Ethnicity  
## Stratification1 StratificationCategoryId1 StratificationID1  
## 1 Hispanic RACE RACEHIS  
## 2 High school graduate EDU EDUHSGRAD  
## 3 $50,000 - $74,999 INC INC5075  
## 4 Data not reported INC INCNR  
## 5 Less than $15,000 INC INCLESS15  
## 6 Hispanic RACE RACEHIS

library(readr)

## Warning: package 'readr' was built under R version 4.4.1

library(gridExtra)

## Warning: package 'gridExtra' was built under R version 4.4.1

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.4.1

library(huxtable)

## Warning: package 'huxtable' was built under R version 4.4.1

##   
## Attaching package: 'huxtable'

## The following object is masked from 'package:ggplot2':  
##   
## theme\_grey

library(dplyr)

## Warning: package 'dplyr' was built under R version 4.4.1

##   
## Attaching package: 'dplyr'

## The following object is masked from 'package:huxtable':  
##   
## add\_rownames

## The following object is masked from 'package:gridExtra':  
##   
## combine

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(broom)

## Warning: package 'broom' was built under R version 4.4.1

library(huxtable)  
library(flextable)

## Warning: package 'flextable' was built under R version 4.4.1

##   
## Attaching package: 'flextable'

## The following objects are masked from 'package:huxtable':  
##   
## align, as\_flextable, bold, font, height, italic, set\_caption,  
## valign, width

###Introduction In this simple analysis we will reorganize the data to create three time series of the united states concerning obesity, physical activity, and fruit and vegetable consumption scores. After analyzing the trends on a national basis we will then go on to compare California and Texas, the two most populous states.

###1. Construct the time series First we simplify the dataset deleting

1. The duplicate variables.
2. The variables not usefull for our analyisis or with too many not available data.

At the end we will remain with this data:

data<-Nutrition\_Physical\_Activity\_and\_Obesity[,-c(2,4,5,6,8,9,10,12,13,14,18,24:33)]

colnames(data)<-c("year","location","topic","value", "low\_conf\_lim","high\_conf\_lim","sample\_size","age","education","gender","income","race")

Now we can check if it s worth keeping the variables with NA. For doing so we build a simple function that counts the percentage of na data for a variable

na\_perc<-function(x){  
 a<-round(sum(is.na(x))/length(x)\*100)  
 paste(a,"%")}  
  
  
a<-na\_perc(data$age)  
b<-na\_perc(data$education)  
c<-na\_perc(data$gender)  
d<-na\_perc(data$income)  
e<-na\_perc(data$race)  
  
tabella <- data.frame( nrow= c("age","education","gender","income","race"),c(a,b,c,d,e))  
colnames(tabella)<-c("percentage of not available data")  
rownames(tabella) <- c("age","education","gender","income","race")  
tabella

|  |  |
| --- | --- |
| **percentage of not available data** |  |
| age | 0 % |
| education | 0 % |
| gender | 0 % |
| income | 0 % |
| race | 0 % |

These variables are for the most part composed of nonavailable data. However, we do not eliminate them right away because since we will be creating subdatasets there may be an adequate number of observations to use them in the future.

In order to create the time series we arrange the data by year

data<-arrange(data,data$year)

We filter for the “US” location

data\_us<- data[data$location== "US", ]

And finally we can filter by topic, creating tree national dataset about:

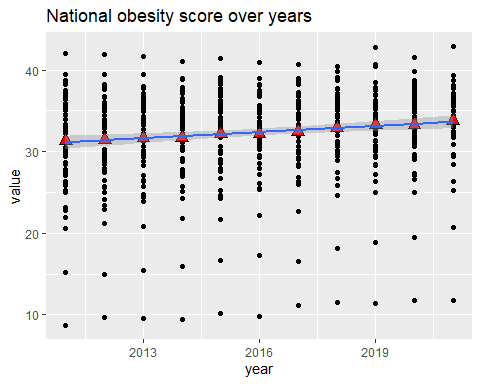
1. Obesity
2. Physical activity
3. Fruit and vegetables consumption

#national obesity dataset over time  
data\_us\_ob<-data\_us[data\_us$topic == "Obesity / Weight Status", ]  
  
#national Physical Activity - Behavior dataset over time  
data\_us\_pa<-data\_us[data\_us$topic == "Physical Activity - Behavior", ]  
  
#national Fruits and Vegetables - Behavior dataset over time  
data\_us\_fv<-data\_us[data\_us$topic == "Fruits and Vegetables - Behavior", ]

###2. National Obesity score over years

ggplot(data = data\_us\_ob, aes(x = year, y= value))+  
 geom\_point()+  
 stat\_summary(  
 geom = "point",  
 fun = "mean",  
 col = "black",  
 size = 3,  
 shape = 24,  
 fill = "red")+  
 geom\_smooth(method = "lm")+  
 ggtitle("National obesity score over years")

## `geom\_smooth()` using formula = 'y ~ x'



We can see that the obesity score has rised over the years, lets check this with a regression line

model<-lm(data= data\_us\_ob , value ~ year)  
huxreg("National obesity score"= model,  
 statistics = FALSE,  
 error\_format = "")

## Warning in huxreg(`National obesity score` = model, statistics = FALSE, : Unrecognized statistics: FALSE  
## Try setting `statistics` explicitly in the call to `huxreg()`

|  |  |
| --- | --- |
|  | National obesity score |
| (Intercept) | -470.836 \*\*\* |
|  |  |
| year | 0.250 \*\*\* |
|  |  |
| \*\*\* p < 0.001; \*\* p < 0.01; \* p < 0.05. | |

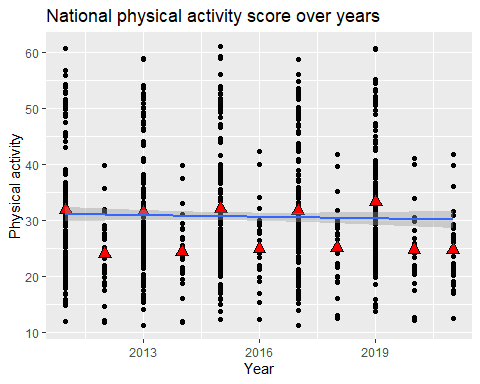
On average adding a year u have an increase of 0.250 on the obesity score that s statistically significant.

It s important to note that we are not going to investigate statistical significance and R squared because these regressions will only serve to give us an idea of the slope of the trend that is more analytical than the visual idea of scatterplots. During the whole analysis we are going to consider a significance level of 90%.

###3. National physical activity score over years

ggplot(data = data\_us\_pa, aes(x = year, y= value))+  
 geom\_point()+  
 stat\_summary(  
 geom = "point",  
 fun = "mean",  
 col = "black",  
 size = 3,  
 shape = 24,  
 fill = "red")+  
 geom\_smooth(method = "lm")+  
 ggtitle("National physical activity score over years")+  
 ylab("Physical activity")+  
 xlab("Year")

## `geom\_smooth()` using formula = 'y ~ x'



Before coming to conclusions about trends in physical activity over the years we note how ratings drop from one year to the next so we need to study what is happening in the data.

a<-sum(data\_us\_pa$year == "2011")  
b<-sum(data\_us\_pa$year == "2012")  
c<-sum(data\_us\_pa$year == "2013")  
d<-sum(data\_us\_pa$year == "2014")  
e<-sum(data\_us\_pa$year == "2015")  
f<-sum(data\_us\_pa$year == "2016")  
g<-sum(data\_us\_pa$year == "2017")  
h<-sum(data\_us\_pa$year == "2018")  
i<-sum(data\_us\_pa$year == "2019")  
j<-sum(data\_us\_pa$year == "2020")  
k<-sum(data\_us\_pa$year == "2021")  
  
tabella <- matrix( c(a,b,c,d,e,f,g,h,i,j,k),nrow = 1)  
colnames(tabella)<-c(2011:2021)  
rownames(tabella) <- c("observations")  
tabella

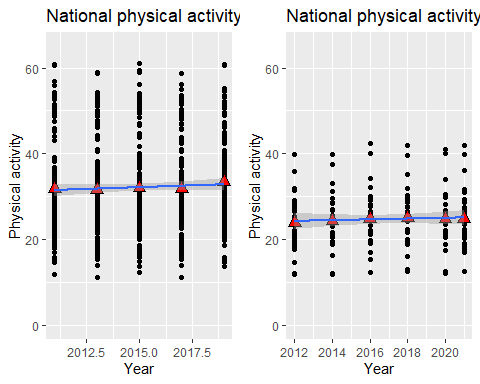
## 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021  
## observations 140 28 140 28 140 28 140 28 140 28 28

Year-to-year observations alternate from 28 to 140 with the last two years being 28.

We then split the data with 28 and 140 observations so that we can see if their trends change.

#create the data subset for samples of 28 and 140 observations  
  
datasub140 <- data\_us\_pa[data\_us\_pa$year %in% c(2011, 2013,2015,2017,2019), ]  
datasub28 <- data\_us\_pa[data\_us\_pa$year %in% c(2012, 2014,2016,2018,2020,2021), ]  
  
  
plot140<-   
 ggplot(data = datasub140, aes(x = year, y= value))+  
 geom\_point()+  
 ylim(0,65)+  
 stat\_summary(  
 geom = "point",  
 fun = "mean",  
 col = "black",  
 size = 3,  
 shape = 24,  
 fill = "red")+  
 geom\_smooth(method = "lm")+  
 ggtitle("National physical activity score over years")+  
 ylab("Physical activity")+  
 xlab("Year")  
  
plot28<-   
 ggplot(data = datasub28, aes(x = year, y= value))+  
 geom\_point()+  
 ylim(0,65)+  
 stat\_summary(  
 geom = "point",  
 fun = "mean",  
 col = "black",  
 size = 3,  
 shape = 24,  
 fill = "red")+  
 geom\_smooth(method = "lm")+  
 ggtitle("National physical activity score over years")+  
 ylab("Physical activity")+  
 xlab("Year")  
  
grid.arrange(plot140,plot28,ncol=2)

## `geom\_smooth()` using formula = 'y ~ x'  
## `geom\_smooth()` using formula = 'y ~ x'



When we have 140 observations your physical activity level is higher on average than when we have 28. The most important thing however is that physical activity has remained stationary in both cases as we can see from the regression (both slope coefficients are low and not significant).

model28<-lm(data= datasub28 , value ~ year)  
model140<-lm(data= datasub140 , value ~ year)  
  
huxreg("model 28 obs" = model28,  
 "model 140 obs" = model140,  
 statistics ="r.squared" ,  
 error\_format = "")

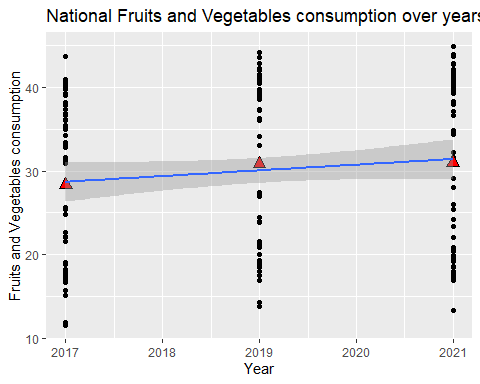
|  |  |  |
| --- | --- | --- |
|  | model 28 obs | model 140 obs |
| (Intercept) | -158.963 | -306.373 |
|  |  |  |
| year | 0.091 | 0.168 |
|  |  |  |
| r.squared | 0.002 | 0.002 |
| \*\*\* p < 0.001; \*\* p < 0.01; \* p < 0.05. | | |

###4. National Fruits and Vegetables consumption over years

For fruit and vegetable consumption we note an increasing trend but the available data only cover 2017, 2019 and 2021.

ggplot(data = data\_us\_fv, aes(x = year, y= value))+  
 geom\_point()+  
 stat\_summary(  
 geom = "point",  
 fun = "mean",  
 col = "black",  
 size = 3,  
 shape = 24,  
 fill = "red")+  
 geom\_smooth(method = "lm")+  
 ggtitle("National Fruits and Vegetables consumption over years")+  
 ylab("Fruits and Vegetables consumption")+  
 xlab("Year")

## `geom\_smooth()` using formula = 'y ~ x'



However, despite the lack of data, we can state that there has not been a conspicuous growth in fruit and vegetable consumption. The slope coefficient indeed is infact not statistically significant.

model\_fv<-lm(data= data\_us\_fv , value ~ year)  
  
huxreg("model fv" = model\_fv,  
 statistics ="r.squared" ,  
 error\_format = "")

|  |  |
| --- | --- |
|  | model fv |
| (Intercept) | -1343.605 |
|  |  |
| year | 0.680 |
|  |  |
| r.squared | 0.013 |
| \*\*\* p < 0.001; \*\* p < 0.01; \* p < 0.05. | |

###5. California and Texas comparison The previous method at the national level can be used to partition data and compare various states. In this example we are going to consider only the data from the 2 most important states by population (California and Texas) so we construct two datasets for the two states.

#california data  
data\_ca<- data[data$location== "CA", ]  
  
#texas data  
data\_tx<- data[data$location== "TX", ]

Then we further partition the dataset for each state by dividing it by obesity, physical activity, and fruit and vegetable consumption. Now we have the time series of the three topics of the analysis and we can investigate them.

#national obesity dataset over time  
data\_ca\_ob<-data\_ca[data\_ca$topic == "Obesity / Weight Status", ]  
#national Physical Activity - Behavior dataset over time  
data\_ca\_pa<-data\_ca[data\_ca$topic == "Physical Activity - Behavior", ]  
#national Fruits and Vegetables - Behavior dataset over time  
data\_ca\_fv<-data\_ca[data\_ca$topic == "Fruits and Vegetables - Behavior", ]  
  
#national obesity dataset over time  
data\_tx\_ob<-data\_tx[data\_tx$topic == "Obesity / Weight Status", ]  
#national Physical Activity - Behavior dataset over time  
data\_tx\_pa<-data\_tx[data\_tx$topic == "Physical Activity - Behavior", ]  
#national Fruits and Vegetables - Behavior dataset over time  
data\_tx\_fv<-data\_tx[data\_tx$topic == "Fruits and Vegetables - Behavior", ]

##5.1 Obesity score comparison

california\_obesity<-ggplot(data = data\_ca\_ob, aes(x = year, y= value))+  
 geom\_point()+  
 stat\_summary(  
 geom = "point",  
 fun = "mean",  
 col = "black",  
 size = 3,  
 shape = 24,  
 fill = "red")+  
 ylim(20,45)+  
 geom\_smooth(method = "lm")+  
 ggtitle("California obesity score over years")+  
 ylab("Obesity Score")+  
 xlab("Year")  
  
texas\_obesity<-ggplot(data = data\_tx\_ob, aes(x = year, y= value))+  
 geom\_point()+  
 stat\_summary(  
 geom = "point",  
 fun = "mean",  
 col = "black",  
 size = 3,  
 shape = 24,  
 fill = "red")+  
 ylim(20,45)+  
 geom\_smooth(method = "lm")+  
 ggtitle("Texas obesity score over years")+  
 ylab("Obesity Score")+  
 xlab("Year")  
  
grid.arrange(california\_obesity,texas\_obesity,ncol=2)

## Warning: Removed 73 rows containing non-finite outside the scale range  
## (`stat\_summary()`).

## `geom\_smooth()` using formula = 'y ~ x'

## Warning: Removed 73 rows containing non-finite outside the scale range  
## (`stat\_smooth()`).

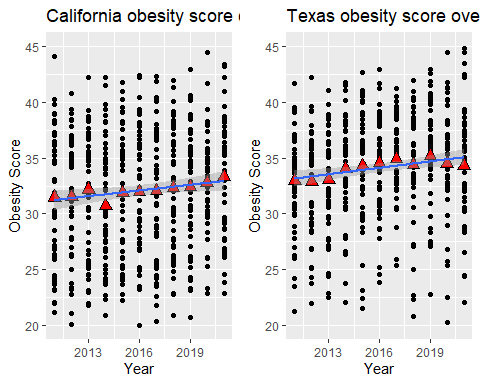
## Warning: Removed 73 rows containing missing values or values outside the scale range  
## (`geom\_point()`).

## Warning: Removed 66 rows containing non-finite outside the scale range  
## (`stat\_summary()`).

## `geom\_smooth()` using formula = 'y ~ x'

## Warning: Removed 66 rows containing non-finite outside the scale range  
## (`stat\_smooth()`).

## Warning: Removed 66 rows containing missing values or values outside the scale range  
## (`geom\_point()`).



Texas has higher levels of obesity, and in both states it has risen. Through the regression table we can identify the linear coefficient to see in which state it grew the fastest.

reg\_ca<-lm(data=data\_ca\_ob, value ~ year )  
reg\_tx<-lm(data=data\_tx\_ob, value ~ year )  
  
  
huxreg("California obesity" = reg\_ca,  
 "Texas obesity" = reg\_tx,  
 statistics = "",  
 error\_format = "")

## Warning in huxreg(`California obesity` = reg\_ca, `Texas obesity` = reg\_tx, : Unrecognized statistics:   
## Try setting `statistics` explicitly in the call to `huxreg()`

|  |  |  |
| --- | --- | --- |
|  | California obesity | Texas obesity |
| (Intercept) | -493.494 \*\* | -450.212 \*\* |
|  |  |  |
| year | 0.260 \*\* | 0.240 \*\* |
|  |  |  |
| \*\*\* p < 0.001; \*\* p < 0.01; \* p < 0.05. | | |

Thanks to the regression we can see that California started from a lower level of obesity but it experienced a more pronounced growth.