R DATABASE PROJECT

Short presentation

Tania Queuche

MSc DMDS

GENERAL INFORMATION

Database name: Nutrition_Physical_Activity_And_Obesity

Source: retrieved on dev.socrata, https://dev.socrata.com/foundry/chronicdata.cdc.gov/hn4x-zwk7, source domain chronicdata.cdc.gov

Date: 2017–2018

Size: 16,384 observations

Description: This dataset includes data on American adult's diet, physical activity, and weight status from Behavioral Risk Factor Surveillance System. This data was originally used for DNPAO's (Division of Nutrition, Physical Activity and Obesity) Data, Trends, and Maps database, which provides national (US) and state specific data on obesity, nutrition, physical activity, and breastfeeding.

Variables: *Age, Education, Gender, Income, Race, Behavioral questions, Percentage of people who answered "yes" to those questions, Location* (55 states)

Type of data: numerical (Data), categorical (Location, Gender, Race, Age, Income), text (Behavioral Questions)

Choice of the dataset

I wanted to choose a dataset related to health or food. Ideally, I want it to be very informative and easy to relate to.

Purpose of the analysis

By analyzing this dataset, I would like to explore the correlations between different factors such as physical activity, gender, race, income, etc. and obesity.

For example, correlations between

Physical activity <> obesity

Location <> obesity

Race <> obesity

Income <> obesity

Ultimately, this kind of analysis can serve informative purposes for the public (spreading awareness about prevalence of obesity) and be used by organizations such as DNPAO in order to lead prevention campaigns where the risks of obesity are the highest.

Feasibility

In order to make sure the analysis was feasible; I verified the consistency of te dataset. After importing it to Excel, I proceeded to:

1) Verify: Is there too much missing data? Is the data easily analyzable? Is there a storytelling opportunity?
2) Classify & Clean on Excel: defining and renaming the variables / columns, removing unnecessary columns
3) Import dataset to R

Variables description

Name	Description	Categories
Age	Categorical data. Age range of the individual at the time surveyed.	Age 18 - 24 25 - 34 35 - 44 45 - 54 55 - 64 65 or older
Education	Categorical data. Highest education level of the individual.	Education College graduate Some college or technical school Less than high school High school graduate
Income	Categorical data. Level of income of the individual, in dollars.	Income Less than $15,000 $15,000 - $24,999 $25,000 - $34,999 $35,000 - $49,999 $50,000 - $74,999 $75,000 or greater
Race	Categorical data. Self-identified race / ethnicity of the individual surveyed.	Race American Indian/Alaska Native Asian Hawaiian/Pacific Islander Hispanic Non-Hispanic Black Non-Hispanic White 2 or more races Other

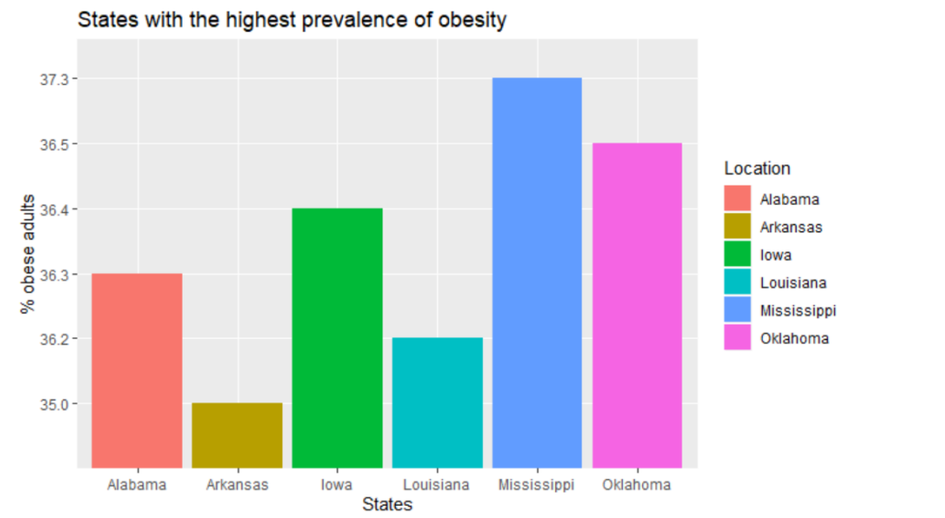
Behavioral Questions

Class	QID	Question
Fruits and Vegetables	Q018	Percent of adults who report consuming fruit less than one time daily
Fruits and Vegetables	Q019	Percent of adults who report consuming vegetables less than one time daily
Obesity / Weight Status	Q036	Percent of adults aged 18 years and older who have obesity
Obesity / Weight Status	Q037	Percent of adults aged 18 years and older who have an overweight classification
Physical Activity	Q043	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 activity (or an equivalent combination)
Physical Activity	Q044	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
Physical Activity	Q045	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)
Physical Activity	Q046	Percent of adults who engage in muscle-strengthening activities on 2 or more days a week
Physical Activity	Q047	Percent of adults who engage in no leisure-time physical activity

Answers to those questions above are numerical data (a percentage). The variable is called Data.

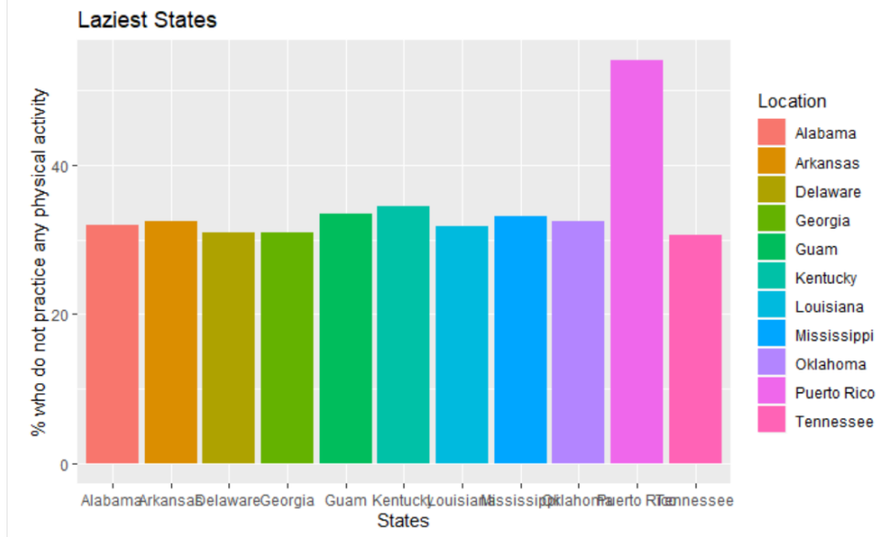
Preliminary Analysis

```r
#obesity by state

R_PROJECT_Nutrition_Physical_Activity_and_Obesity %>%
  filter(Value=="General",                    # filtering only "General" because we
want the values for every population category!
         Question_ID=="Q036",
         Data>= 35) %>%                       # taking only the states for which X%
of the population is obese
  arrange(desc(Data)) %>%

# plotting a graph for the worse states
ggplot(aes(x=Location, y=Data , fill=Location)) + geom_col() +
labs(title="States with the highest prevalence of obesity",
     y="% obese adults",
     x="States")
```



Graph representing the 6 states with the highest rates of obesity among adults.

```r
# laziest states
ob_data %>%
  filter(Value=="General",
         Question_ID=="Q047",
         Data >= 30) %>%
  arrange(desc(Data)) %>%

# plotting a graph for the worse states
ggplot(aes(x=Location, y=Data , fill=Location)) + geom_col() +
labs(title="Laziest States",
     y="% who do not practice any physical activity",
     x="States")
```



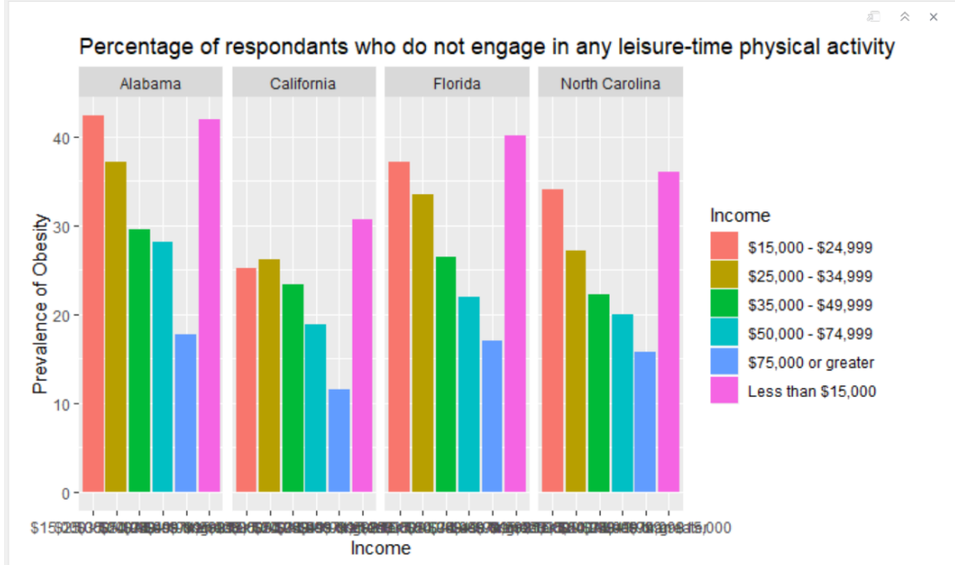
45% of the laziest states (where the highest percentage of the population does not practice any physical activity) are among the top 6 states for obesity. Coincidence?

```{r}
ob_data%>%
  group_by(Income)%>%
filter(Value=="Specific",
       Question_ID=="Q036",
       Location%in% c("California","North Carolina","Florida",
"Alabama"),
       Income!="NA",
       Income!="Data not reported")%>%


ggplot(aes(x=Location, y=Data, color=Income)) + geom_point(size=7,
alpha=0.5) + labs(title="Prevalence of Obesity by Income",
    x="State",
    y="Prevalence of Obesity",
    theme(text=element_text(size=14))
)
```

Interestingly, for all states represented on this graph, the highest point is **pink** (less than $15k) and the lowest one is **blue** ($75k or greater); it is reasonable to guess that obesity is inversely proportional to income.

There seems to be a correlation between income and obesity.

```{r}

ob_data%>%

filter(Value=="Specific",
       Question_ID=="Q047",
       Location%in% c("California","North Carolina","Florida", "Alabama"),
       Income!="NA",
       Income!="Data not reported")%>%

  ggplot(aes(x=Income, y=Data, fill=Income))+geom_col()+facet_grid(~Location) +
labs(title="Percentage of respondants who do not engage in any leisure-time
physical activity",
    y="Prevalence of Obesity",
    theme(text=element_text(size=14),
      axis.text.x=element_text(color="#FFFFFF"))
    )
```



The lower the income, the less people are likely to practice leisure-time physical activities. Could this be related to the high obesity rates among low-income?

Next steps

Improving data visualisation:

- ➔ **Define order for categorical data: order Income, Age, Education**
- ➔ **Change the color scheme: for categories, use a gradients (for example** *hp+scale_fill_gradient(low="yellow", high="red")*

Going further in the analysis:

- ➔ **Identify correlations using ggplot and gganimate:**

- *How does behaviour, such as consumption of fruits & vegetables and practice of physical activity, influence obesity?* I will represent the correlations thanks to scatter plots or geom_point plots .

- *Are certain groups more prone to Obesity? Are we all equal concerning Obesity?* I will visualize the behaviours by group thanks to histograms (similar to the Prevalence of Obesity by Income).

- *After the previous analysis, I expect to have a few hypothesis about which groups are most prone to obesity and for which reasons. Do all Income/Gender/Age/Education/Race groups have the same behaviour? Is one particular group often practising risky behaviour?* I will visualize the behaviours by group thanks to histograms, such as the histogram on the previous page).

Creating an application (?):

- Creating a sidebar with checkboxGroupInput()
- Creating tabs corresponding to each question using tabBox()