**Halloween is coming..!**

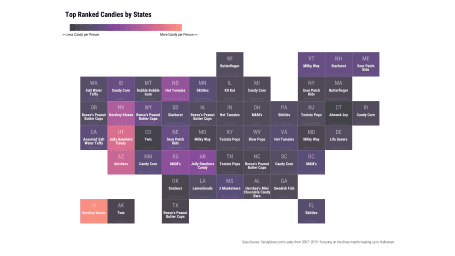
Halloween is just around the corner, I am still trying to decide which candies to purchase this year for trick-or-treaters.

Initially I was looking for data sets maybe comparing American chocolate bars vs Canadian chocolate bars possibly with sugar contents or lists of ingredients. I am really curious why there are big differences in candy between Canada and US, but for now I couldn’t find them instead I came across below data sets.

Data itself is actually bit outdated, since it’s data from candy sales in 2007-2015. I’m not even sure if these popular candies changes every year or not. Now I’m curious to find data sets that are more recent…  
[CandyStore.com’s sales from 2007–2015—focusing on the three months leading up to Halloween](https://public.opendatasoft.com/explore/embed/dataset/state-by-state-favorite-halloween-candy/table/)

Since data set contains data at state level for US, I decided I’ll practice plotting the data on grid map, I’ve been wanting to try. To colour the map, I’ve decided instead of colouring by volume of candy purchased (in pounds), I’ve gathered population by state in order to figure out ratio of candy per person.

Looks like Hawaiian eats lots of Hershey Kisses (or in general, maybe there are more sweet tooth in Hawaii!). Utah, Nevada and Arizona also seems to maybe either consumer more sweets or possibly top ranked chocolates are purchased in volume!



It would’ve been fun to gather all candy images and place them in the grid too…!

Below are the steps I took to visualize the data.

**Importing Data & Tidying Up**

I’ve imported data to dataframe using [jsonlite](https://cran.r-project.org/web/packages/jsonlite/index.html) package using fromJSON function from [OpenDataSoft site](https://public.opendatasoft.com/explore/?sort=modified). There seems to be lots of other interesting data sets!

I’ve also harvested population by state from Wikipedia using [rvest](https://cran.r-project.org/web/packages/rvest/) package.

library(tidyverse) ## basically tidyverse is needed for everything :)

library(jsonlite) ## to read data in json format

library(rvest) ## to scrape website

library(hrbrthemes) ## I love this themes!

## read this: https://cran.r-project.org/web/packages/hrbrthemes/vignettes/why\_hrbrthemes.html

library(geofacet) ## I'll use the grid in this package to create

library(ggrepel) ## so that texts on ggplot don't overlap!

library(janitor) ## My recent favourite

## Download Candy Data from Opendata Soft site.

candy\_data <- jsonlite::fromJSON("https://public.opendatasoft.com/api/records/1.0/search/?rows=51&start=0&fields=name,top\_candy,top\_candy\_pounds,2nd\_place,2nd\_place\_pounds,3rd\_place,3rd\_place\_pounds&dataset=state-by-state-favorite-halloween-candy&timezone=America%2FLos\_Angeles")

## I'll call this candy\_wide, because data is in wide format.

candy\_wide <-candy\_data$records$fields

names(candy\_wide) <- c("state","rank\_3","pound\_2","rank\_2","pound\_3","rank\_1","pound\_1")

## Might be interesting to see Volume per Capita, so get population data from Wiki.

state\_pop\_html <-

read\_html("https://simple.wikipedia.org/wiki/List\_of\_U.S.\_states\_by\_population") %>%

html\_table(2)

state\_pop\_html <- state\_pop\_html[[2]]

## I'm only really interested in population, so trim down the table bit.

state\_pop <- state\_pop\_html %>% select(state=3, pop\_2016=4)

## I want population to be numeric value instead of character

state\_pop <- state\_pop %>% mutate(pop\_2016 = as.numeric(str\_remove\_all(pop\_2016,",")))

**Transforming Data from Wide to Long format**

### Data Transformaiton (Wide to Long)

## Data is in wide format, nicer to look at but I want to convert it to long format!

candy\_rank\_long <- candy\_wide %>%

select(state, rank\_1, rank\_2, rank\_3) %>%

gather(ranking, chocolate, -state) %>%

mutate(ranking = as.integer(str\_remove(ranking,"rank\_")))

candy\_pound\_long <- candy\_wide %>%

select(state, pound\_1, pound\_2, pound\_3) %>%

gather(ranking, volume, -state) %>%

mutate(ranking = as.integer(str\_remove(ranking,"pound\_")))

candy\_long <- candy\_rank\_long %>%

inner\_join(candy\_pound\_long) %>%

left\_join(state\_pop)

candy\_long <-candy\_long %>% mutate(volume\_per\_capita = volume/pop\_2016)

## Just creating state table

state\_df <- tibble(

state = c(state.name,"District of Columbia"),

state\_div = c(as.character(state.division),"South Atlantic"),

state\_abb = c(state.abb, "DC")

)

candy\_long <- candy\_long %>%

left\_join(state\_df)

**Overall Popular Candies in United States**

If candy was ranked #1 in state, I gave gold colour, if #2, then silver, #3 then bronze. Sort of like Olympic, I’ve counted how many metals each candies have gotten to decide most popular candy in US.

There were total of 27 different candies in this data sets, most popular candies are M&M followed by Skittles! I thought it was also interesting that Life Savers were top ranked candy in Delaware, but it did not appear in any other state, and similarly for Swedish Fish in Georgia.

There’s Assorted Salt Water Taffy and Salt Water Taffy… I wasn’t sure if they are same candies…!

## using janitor package

candy\_list <- candy\_long %>% tabyl(chocolate, ranking) %>%

adorn\_totals("col") %>%

arrange(-Total,-`1`,-`2`,-`3`) ## if candy was listed same # of times, then I want to make sure that better ranked candies are ranked higher.

#show\_col(c("#A77044", "#A7A7AD","#D6AF36")) # bronze, silber, gold

candy\_long %>%

mutate(chocolate = fct\_relevel(chocolate, candy\_list$chocolate)) %>%

count(chocolate, ranking, state, state\_abb) %>%

ggplot(aes(x=fct\_rev(chocolate),y=n)) +

geom\_col(aes(fill=fct\_rev((factor(ranking)))), colour="#ffffff30") +

geom\_text(aes(label=state\_abb), position="stack", colour="#ffffff",

size=5, hjust=1, family=font\_rc, fontface="bold") +

theme\_ipsum\_rc() +

coord\_flip() +

scale\_fill\_manual(values=c("#A77044", "#A7A7AD","#D6AF36"),

name="ranking within state",

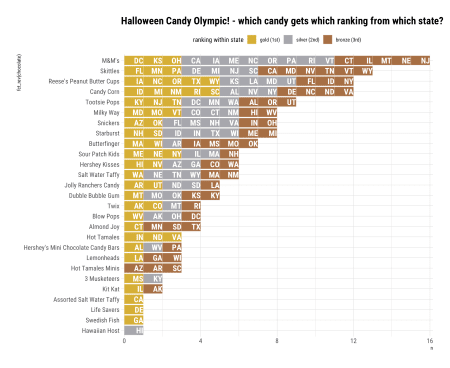
labels=c("bronze (3rd)", "silver (2nd)","gold (1st)")) +

scale\_y\_comma() +

labs(title="Halloween Candy Olympic! - which candy gets which ranking from which state?") +

theme(legend.position="top", legend.direction = "horizontal") +

guides(fill = guide\_legend(reverse = TRUE))



**Grid State Map**

I recently discovered [geofacet](https://cran.r-project.org/web/packages/geofacet/index.html) package. You can create with this package looks super fun! I thought it was really neat that I can create my own grid using “Geo Grid Designer”!!

#get\_grid\_names() --> you can see what grids are available.

## I can preview the grid first

#grid\_design(us\_state\_grid1)

## Going to use US state map grid in geofacet package

head(us\_state\_grid1) %>% knitr::kable()

| **row** | **col** | **code** | **name** |
| --- | --- | --- | --- |
| 6 | 7 | AL | Alabama |
| 7 | 2 | AK | Alaska |
| 5 | 2 | AZ | Arizona |
| 5 | 5 | AR | Arkansas |
| 4 | 1 | CA | California |
| 4 | 3 | CO | Colorado |

## Join Candy data & grid so I can create grid map

candy\_long\_with\_coord <-

candy\_long %>%

left\_join(us\_state\_grid1, by=c("state\_abb" = "code"))

## Replace every other space with new line

## some chocolate names are little too long to lists in one line...

space\_to\_newline = function(x) gsub("([^ ]+ [^ ]+) ","\\1\n", x)

candy\_long\_with\_coord <- candy\_long\_with\_coord %>%

mutate(chocolate\_label = map\_chr(chocolate, space\_to\_newline),

facet\_label = case\_when(ranking==1 ~ "Top Ranked Candies by State",

ranking==2 ~ "2nd Popular Candies by States",

ranking==3 ~ "3rd Popular Candies by States"))

candy\_long\_with\_coord %>%

#filter(ranking==1) %>%

ggplot(aes(x=col, y=row)) +

geom\_tile(color="#ffffff90", aes(fill=volume\_per\_capita)) +

#geom\_text(aes(label=state), size=3.5, color="#ffffff20", vjust=4, family=font\_rc\_light) +

geom\_text(aes(label=state\_abb),size=5, family=font\_rc, vjust=-1, color="#ffffff90") +

geom\_text(aes(label=chocolate\_label), family=font\_rc, size=3.5, lineheight=0.8, color="#ffffff", vjust=1, fontface="bold") +

scale\_y\_reverse(breaks=NULL) +

scale\_x\_continuous(breaks=NULL) +

theme\_ipsum\_rc(base\_family="Roboto Condensed") +

scale\_fill\_viridis\_c(option="magma", alpha=0.8, end=0.7,

name="",

breaks=fivenum(candy\_long$volume\_per\_capita),

labels=c("","","<< Less Candy per Person","","More Candy per Person >>")) +

facet\_wrap(~fct\_reorder(facet\_label, ranking), ncol=1) +

theme(legend.position = "top",

legend.key.width = unit(2, "cm"),

legend.justification = "left") +

labs(title="Top Ranked Candies by States",

caption="Data Source: CandyStore.com's sales from 2007–2015—focusing on the three months leading up to Halloween", x="", y="") +

coord\_fixed(ratio=0.9)

