Lab 6 - La Quinta is Spanish for next to Denny's, Pt. 1

Visualizing spatial data

https://datasciencebox.org/

Have you ever taken a road trip in the US and thought to yourself "I wonder what La Quinta means". Well, the late comedian Mitch Hedberg (https://en.wikipedia.org/wiki/Mitch_Hedberg) thinks it's Spanish for *next to Denny's*.

If you're not familiar with these two establishments, Denny's is a casual diner chain that is open 24 hours and La Quinta Inn and Suites is a hotel chain.

These two establishments tend to be clustered together, or at least this observation is a joke made famous by Mitch Hedberg. In this lab we explore the validity of this joke and along the way learn some more data wrangling and tips for visualizing spatial data.

The inspiration for this lab comes from a blog post by John Reiser on his *new jersey geographer* blog. You can read that analysis http://njgeo.org/2014/01/30/mitch-hedberg-and-gis/. Reiser's blog post focuses on scraping data from Denny's and La Quinta Inn and Suites websites using Python. In this lab we focus on visualization and analysis of these data. However note that the data scraping was also done in R, and we we will discuss web scraping using R later in the course. But for now we focus on the data that has already been scraped and tidied for you.

Learning goals

- Visualising spatial data
- Joining data frames

#What is Spatial Data?

Spatial data, also known as geospatial data, is a term used to describe any data related to or containing information about a specific location on the Earth's surface.

Getting started

As usual, clone Lab6 files to your github account by clicking on plus (+) on the top-right corner of your account. Select Import repository and add the link https://github.com/musabu/Lab6.git. Name the repository Lab6, make it private and click Begin import.

Click the name of the created repo and copy the link to the repo by clicking on the green button named Code. Go to your RStudio cloud profile. Select DSC200 -> Click on the down arrow next to the New Project button -> Select New Project from Git Repository -> Paste the link you copied for Lab6 git repository copied from your account (**NOT FROM musabu/Lab6.git**) and click OK. Open the file Lab6-solution.Rmd to answer the lab exercises.

Ensure you add me (using musabu username) as a collaborator so that I can review your work later. To do this go to the repo page, click on Settings -> Manage Access.

Warm up

Before we introduce the data, let's warm up with some simple exercises.

- Update the YAML, changing the author name to your name, and **knit** the document.
- Commit your changes with a meaningful commit message.
- Push your changes to GitHub.
- Go to your repo on GitHub and confirm that your changes are visible in your Rmd **and** md files. If anything is missing, commit and push again.

Packages

We'll use the **tidyverse** package for much of the data wrangling and visualisation and the data lives in the **dsbox** package. These packages are already installed for you. You can load them by running the following in your Console:

```
library(tidyverse)
library(dsbox)
```

Note: dsbox is not yet on CRAN. For now, you can install it from GitHub with the following commands on the terminal

```
install.packages("devtools")
devtools::install_github("rstudio-education/dsbox")
```

Data

The datasets we'll use are called `dennys` and `laquinta` from the **dsbox** package. Note that these data were scraped from https://locations.dennys.com/ and https://www.lq.com/en/findandbook/hotel-listings.html, respectively.

The datasets we'll use are called `dennys` and `laquinta` from the **dsbox** package. Since the datasets are distributed with the package, we don't need to load them separately; they become available to us when we load the package. You can find out more about the datasets by inspecting their documentation, which you can access by running `?dennys` and `?laquinta` in the Console or using the Help menu in RStudio to search for `dennys` or `laquinta`. You can also find this information https://rstudio-education.github.io/dsbox/reference/dennys.html and https://rstudio-education.github.io/dsbox/reference/laquinta.html.

To help with our analysis we will also use a dataset on US states, which is located in your repository's `data` folder.

```
states <- read_csv("data/states.csv")
```

Each observation in this dataset represents a state, including DC. Along with the name of the state we have the two-letter abbreviation and we have the geographic area of the state (in square miles).

Exercises

1. What are the dimensions of the Denny's dataset? (Hint: Use inline R code and functions like `nrow` and `ncol` to compose your answer.) What does each row in the dataset represent? What are the variables?
2. What are the dimensions of the La Quinta's dataset? What does each row in the dataset represent? What are the variables?

Knit, *commit, and push your changes to GitHub with an appropriate commit message. Make sure to commit and push all changed files so that your Git pane is cleared up afterwards.*

We would like to limit our analysis to Denny's and La Quinta locations in the United States.

3. Add a country variable to the Denny's and La Quinta's datasets and set all observations equal to `"United States"`. Remember, you can use the `mutate` function for adding a variable. Make sure to save the result of this as `dn` and lq, respectively, so that the stored data frame contains the new variable going forward.

We don't need to tell R how many times to repeat the character string "United States" to fill in the data for all observations, R takes care of that automatically.

```
dn <- dennys %>%
  mutate(country = "United States")
```

Add the code below for La Quinta's dataset

Knit, *commit, and push your changes to GitHub with an appropriate commit message. Make sure to commit and push all changed files so that your Git pane is cleared up afterwards.*

4. Which states have the most and fewest Denny's locations? What about La Quinta? Is this surprising? Why or why not?

5. Next, let's calculate which states have the most Denny's and La Quinta's locations *per thousand square miles*. This requires *joinining* information from the frequency tables you created in Exercise 3 with information from the `states` data frame.

First, we count how many observations are in each state, which will give us a data frame with two variables: `state` and `n`.

Then, we join this data frame with the `states` data frame. However note that the variables in the `states` data frame that has the two-letter abbreviations is called `abbreviation`.

So when we're joining the two data frames we specify that the `state` variable from the Denny's data should be matched `by` the `abbreviation` variable from the `states` data:

```
dn %>%
  count(state) %>%
  inner_join(states, by = c("state" = "abbreviation"))
```

So to answer the question - Denny's and La Quinta's locations *per thousand square miles*. We put the two datasets together into a single data frame.

However before we do so, we need to add an identifier variable. We'll call this `establishment` and set the value to `"Denny's"` and `"La Quinta"` for the `dn` and `lq` data frames, respectively.

```
dn <- dn %>%
  mutate(establishment = "Denny's")
lq <- lq %>%
  mutate(establishment = "La Quinta")
```

Since the two data frames have the same columns, we can easily bind them with the `bind_rows` function:

```
dn_lq <- bind_rows(dn, lq)
```

We can plot the locations of the two establishments using a scatter plot, and color the points by the establishment type. Note that the latitude is plotted on the x-axis and the longitude on the y-axis.

```
ggplot(dn_lq, mapping = aes(x = longitude, y = latitude, color = establishment)) +
  geom_point()
```

Knit, *commit, and push your changes to GitHub with an appropriate commit message. Make sure to commit and push all changed files so that your Git pane is cleared up afterwards and review the md document on GitHub to make sure you're happy with the final state of your work.*