Problem Set 1

whomsoever

Introduction

This week's problem set will guide you through some simple tools for analyzing data. We've provided four datasets for you to choose from. You may also work on your own–but please consult me first.

This document was produced by R using R Markdown. The advantage of using R Markdown is you can present your code and text annotations in the same document. Don't worry if you aren't yet accustomed to R–the coding parts of this problem set are mostly expositional. We encourage you to use R Markdown for this problem set, but if you'd rather use a different environment, e.g. Jupyter notebooks (etc), you are welcome to.

For tips on R Markdown formatting and syntax, check out the course notes (chapter 1) or the cheatsheet we've provided.

To compile a R Markdown file (i.e. create the output document), click 'Knit'. For pdf output, you'll need to have some form of TeX installed on your computer–https://www.latex-project.org/get/. Alternatively you can use html output, by specifying `html_document` in the YAML metadata at the very top of the document, which does not require TeX.

Any time you are confused about a command in R, use `help()` to bring up the relevant documentation. If you get an error message you don't understand, use google–it's very likely someone else has encountered the same error and written about it on Stack Overflow or the like.

The packages used in this problem set are:

- `tidyverse` (is a set of eight packages)
- `stats`

Make sure these are loaded in the setup chunk using the `library()` command. If you get the error message "there is no package called 'XXX' ", you need to first install the package using `install.packages('XXX')`, then load it into the session using `library()`.

I – Warmup

This section is optional, but if you are new to R you may find it worthwhile.

II – The NYC Heatwave Dataset

For the remainder problem set you will use the NYC Heatwave dataset (filename 'nyc-heatwave.csv'). It contains data on the temperature and urban characteristics of 1021 locations in the city during a heatwave.

Your first order of action is loading the data into your R session. First download the data file and save it somewhere convenient on your computer (e.g. the same directory your code file is in). There are many functions you can use to load data into R. For csv data you can use `read.csv()`. In the argument you must specify the file path to your data. If you don't know how to find a file path, give it a google.

Note, if your data is in the same directory as your code file, you don't need to specify the file path–simply the file name will do. If you are having issues, you can check the current working directory of your R session

using the `getwd()` command. You can also use `setwd()` to manually specify a working directory for your R session.

1. In the chunk below load the heatwave data into R. If your data is in the same directory as your code file, the sample code below should work–simply uncomment it and run it. Note, "mydata" is a terrible name for a data frame–try to use something more relevant.

```
mydata <- read.csv('../data/nyc-heatwave.csv')
```

Great. Your dataset has now been loaded and assigned to an object.

There are two ways to view data you have loaded into R:

- use the `View()` command to view the entire dataset in a separate window
- use the `head()` or `tail()` commands to view the first or last few rows of the dataset

Don't use the `View()` command for massive datasets as it is very memory intensive and might crash your computer.

Here is a brief description of the variables and what exactly they are measuring:

- **area**–geographical area of the recorded location
- **temperature**–the temperature of the recorded location in degrees Farenheit
- **vegetation**–a measure of the relative concentration of vegetation (i.e. greenery) in the recorded area. It is measured in NDVI (normalized difference vegetation index), a number between -1 and 1. 0 indicates no vegetation, 1 indicates maximum vegetation, and negative values indicate water.
- **albedo**–a measure of the solar reflectivity of buildings in the recorded area. E.g. an albedo of 25% means buildings reflect on average 25% of solar radiation.
- **building height**–the average height of buildings in the recorded area. Its units are storeys.

Now run the following checks on your dataset:

2. Check the class of your dataset using `class()`:

3. Print the column names of your dataset using `colnames()`:

4. Check the data type of each column using `class()`:

5. Are there categorical variables in the data? If so, which ones? What are the levels/categories of the categorical variable(s)? Use `summary()` or `unique()` on the variable to pull up its levels.

6. Print the dimensions of your dataset (number of rows and columns) using `dim()`:

III – Visualizing Data

Create histograms

Create boxplots across categorical variables

IV – Summarizing Data

The mean is a measure of central tendency–it gives the central or typical value of a distribution. The numerical mean of a set of observations is:

$$\bar{x} = \frac{1}{n} \sum_{i}^{n} x_i$$

where x_i denotes an individual observation in the set, and n is the number of observations.

In R you can simply use `mean()` to calculate the mean of a vector array of data.

10. Calculate the mean temperature recorded in the dataset.

The median is another measure of central tendency. It is defined

$$median =$$

11. Calculate the median temperature recorded in the dataset. Is this value close to the mean?
12. Calculate the mean and median albedo recorded in the dataset. Are the values similar?
13. Create a histogram of the variable albedo. What do you see?
14. Using this histogram, comment on whether the mean or median is a better measure of central tendency for albedo.

The standard deviation is a measure of spread. It gives the average distance of each observation from the sample mean. It is defined

$$sd =$$

Aggregate

V – Intervals

compute a confidence interval for a mean

VI – Tests

do a hypothesis test for a difference in sample means

VII – Relationships between variables

association with scatterplots

correlation with cor

correlation matrix

is correlation causation?

VIII – Basic Linear Models

make some basic linear models

simple linear model

multiple regression

building a prediction

interpret coefficients

categorical predictors, interpreting coefficients