

CS636 Homework 1

- Submit electronic copy in Canvas

For the following questions, please use R commands to find solutions when applicable. Please use Jupyter Notebook and submit Notebook File (.ipynb) of your notebook, including all code and results.

1.20 The built-in data set `islands` contains the size of the world's land masses that exceed 10,000 square miles. Use `sort()` with the argument `decreasing=TRUE` to find the seven largest land masses.

For Example, the expected solution is

```
> sort(islands, decreasing=TRUE)[1:7]
      Asia      Africa North America South America  Antarctica      Europe      Australia
16988    11506     9390      6795      5500      3745      2968
```

1.21 Load the data set `primes` (`UsingR`). This is the set of prime numbers in `[1,2003]`. How many are there? How many in the range `[1,100]`? `[100,1000]`?

1.22 Load the data set `primes` (`UsingR`). We wish to find all the twin primes. These are numbers p and $p+2$, where both are prime.

1. Explain what `primes[-1]` returns.
2. If you set `n=length(primes)`, explain what `primes[-n]` returns.
3. Why might `primes[-1]`—`primes[-n]` give clues as to what the twin primes are? How many twin primes are there in the data set?

1.23 For the data set `treering`, which contains tree-ring widths in dimension-less units, use an R function to answer the following:

1. How many observations are there?
2. Find the smallest observation.
3. Find the largest observation.
4. How many are bigger than 1.5?

1.24 The data set `mandms` (`UsingR`) contains the targeted color distribution in a bag of M&Ms as percentages for various types of packaging. Answer these questions.

1. Which packaging is missing one of the six colors?
2. Which types of packaging have an equal distribution of colors?
3. Which packaging has a single color that is more likely than all the others? What color is this?

1.25 The `times` variable in the data set `nym. 2002` (UsingR) contains the time to finish for several participants in the 2002 New York City Marathon. Answer these questions.

1. How many times are stored in the data set?
2. What was the fastest time in minutes? Convert this into hours and minutes using R.
3. What was the slowest time in minutes? Convert this into hours and minutes using R.

1.26 For the data set `rivers`, which is the longest river? The shortest?

1.27 The data set `uspop` contains decade-by-decade population figures for the United States from 1790 to 1970.

1. Use `names()` and `seq()` to add the year names to the data vector.
2. Use `diff()` to find the inter-decade differences. Which decade had the greatest increase?
3. Explain why you could reasonably expect that the difference will always increase with each decade. Is this the case with the data?