STAT 33A Homework 2

CJ HINES (3034590053)

Sep 24, 2020

This homework is due **Sep 24, 2020** by 11:59pm PT.

Homeworks are graded for correctness.

As you work, write your answers in this notebook. Answer questions with complete sentences, and put code in code chunks. You can make as many new code chunks as you like.

Please do not delete the exercises already in this notebook, because it may interfere with our grading tools.

You need to submit your work in two places:

- Submit this Rmd file with your edits on bCourses.
- Knit and submit the generated PDF file on Gradescope.

Exercise 1

Download the Significant Earthquake Database, `quakes.txt`, from the bCourse. Pay attention to where you save the file.

Read about the `stringsAsFactors` parameter in `?read.table`.

1. Use R to load the file into a variable named `quake`. Make sure that strings are not converted into factors.

   Hint: Don't use `read.table()` for this step. Use one of its relatives instead.

2. How many rows and columns are in the dataset?

3. What are the classes of the columns?

YOUR ANSWER GOES HERE:

1.

```
quake = read.delim("quakes.txt")
```

2. 6210 Rows. 47 Columns.

```
numrows = nrow(quake)
numcolumns = ncol(quake)
```

3.

```
col = colnames(quake)
class(col)
```

```
## [1] "character"
```

Exercise 2

The Significant Earthquake Database is originally from this page.

That page links to "Event Variable Definitions", which are definitions of some of the columns in the data set.

When you load a data set from a delimited file, you'll often need to convert some of the columns to the correct class. Some common conversions are:

- character to factor (especially for R >= 4.0 or `stringsAsFactors = FALSE`)
- factor to character (especially for R < 4.0 or `stringsAsFactors = TRUE`)
- integer to factor
- character to numeric
- character to date
- character to logical

R provides `as.` functions to convert to various classes. For instance, `as.character()` converts its argument into a character vector.

1. Categorical data take on a limited number of possible values (categories) and are usually qualitative rather than quantitative.

   Based on the column definitions and the data set itself, which columns are categorical?

   Hint: You can use the `table()` function to see how many different values are in a column.

2. Convert the columns you listed in 2.1 into factors. For each column, you'll need to use `$` both to get and to set the column.

YOUR ANSWER GOES HERE:

1. Categorical Columns: COUNTRY, FLAG_TSUNAMI, LOCATION_NAME, STATE

2.

```
Country = as.factor(quake$COUNTRY)
FlagTsunami = as.factor(quake$FLAG_TSUNAM)
Location = as.factor(quake$LOCATION_NAME)
State = as.factor(quake$STATE)
```

Exercise 3

1. Which country had the most significant earthquakes?

   Hint: When it comes to sorting and subsets, tables are a lot like vectors.

2. Which country (among those recorded in the data set) had the fewest significant earthquakes?

3. What's the location of the earthquake that damaged the most houses, and what year did it happen in?

YOUR ANSWER GOES HERE:

1. China

```
t = table(quake$COUNTRY)
MostCountry = which.max(t)
names(MostCountry)
```

```
## [1] "CHINA"
```

2. Barbados

```
MinCountry = which.min(t)
names(MinCountry)
```

```
## [1] "BARBADOS"
```

3. China 2008

```
MaxDamagedHouses = which.max(quake$TOTAL_HOUSES_DAMAGED)
quake$COUNTRY[MaxDamagedHouses]
```

```
## [1] "CHINA"
```

```
quake$YEAR[MaxDamagedHouses]
```

```
## [1] 2008
```

Exercise 4

1. Compute the median latitude and longitude for the data set. This gives an idea of where the geographical "center" of the earthquakes is.
2. What country is the median point from 2.1 in?

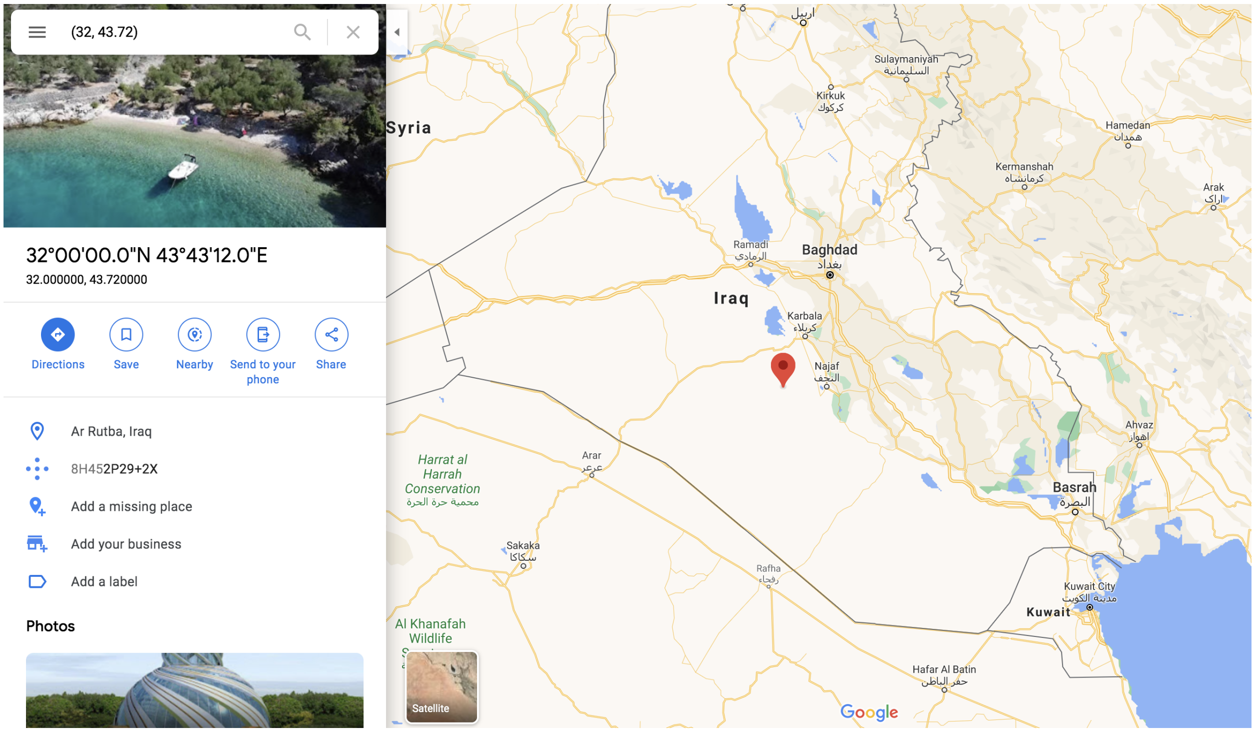
   Use Google Maps or another map service to check. The order of the coordinates should be `(latitude, longitude)`. Take a screenshot of the point in the country and use the Markdown image syntax `![title](/path/to/file)` to insert the screenshot in your notebook.

YOUR ANSWER GOES HERE:

1. (32, 43.72)

```
lat = median(quake$LATITUDE, na.rm = TRUE)
long = median(quake$LONGITUDE, na.rm = TRUE)
```

2. Iraq. Screenshot below.



(32, 43.72)

32°00'00.0"N 43°43'12.0"E

32.000000, 43.720000

Ar Rutba, Iraq

8H452P29+2X

Figure 1: Screenshot

Exercise 5

1. Are there any earthquakes where the country is missing? How can you tell?
2. What does the empty string "" mean in the FLAG_TSUNAMI column?

   Are there any earthquakes where the entry for the tsunami flag is missing?
3. When passed two equal-length vectors as arguments, the `table()` function makes a two-way table.

   You can subset a two-way table the same way you would subset a matrix.

   Which three countries have the highest finite ratio of tsunami relative to non-tsunami?

YOUR ANSWER GOES HERE:

1. No. There are 0 NA values.

```
missingcountry = table(quake$COUNTRY, useNA = "always")
```

2. The empty string in FLAG_TSUNAMI means that the earthquake did not cause a tsunami. Yes there are 4371 entries missing a tsunami flag.

```
tsu = table(quake$FLAG_TSUNAMI)
```

3. SAMOA, NEW CALEDONIA, and JAPAN.

```r
ratio = table(quake$COUNTRY, quake$FLAG_TSUNAMI)
tsunottsu = ratio[,2]/ratio[,1]
s = sort(tsunottsu, decreasing=TRUE)
s[7:9]
```

```
##         SAMOA NEW CALEDONIA         JAPAN
##      7.000000      4.000000      2.530435
```