R Programming

Take Home MidTerm

Submit your solution to the following as an R Notebook, along with the converted csv file, on Camino.

PART 1

1. Load the titanic.csv data

2. Get rid of the following columns:
3. 'Name', 'Ticket', 'PassengerId', 'Cabin'

4. Replace the ‘Sex column with a binary column for whether the passenger is ‘Male’

5. Show how many NA’s are in each column

6. Which Column has the most NA’s? How many does it have?

7. Create a bar chart to show NA counts by column

8. Show/plot the distribution of ages for each gender. Is the average (mean) age of Males higher? Is the median higher?

9. Which port of origin (embarked column) had the highest average fare? Show/plot your results.

10. Is there a statistically significant difference in average fares between the genders?

PART 2

Data Description
c: country cy: city hh: domain r: url redirect tz: time zone Full documentation here: https://dev.bitly.com/nsq.html

1. Load the bitly data from data.gov as a data frame.

2. How many records are there?

3. In the City column (‘cy’), how many are NA?

4. How many countries are present in this data? Which country (‘c’ column) has the most records? How many does it have?

5. How many records are from Russia?
6. How many records do NOT list USA
7. how many records have 'America' in the timezone, but Country not in US?

8. How many records point to cia.gov?

9. Fill empty records with “UNKNOWN” and NA’s with “MISSING”. Create a new csv. (Submit this csv with your R Notebook.)

10. Show the top 10 timezones (‘tz’ column) in a bar chart, with a legend, and properly labeled x,y axes and title.