PS1: Data cleaning and data-based discussion

Pooja Sadarangani

2022-10-15

Basic R Programming

Question 1: Write a function that takes in time in the form of HHMM (hours-minutes) as a number and returns it as HH.HH (hours, fractions of hours) as a number. For instance, it should convert 730 (7h 30m) into 7.5 (7.5 hrs) and 1245 into 12.75. Assume the argument is passed as a number, and it should return a number, i.e. not print it but return.Test it demonstrating that 730 and 1245 are converted correctly.

```
## [1] 7.5
```

```
## [1] 12.75
```

Question 2: Use a for-loop to extract only positive numbers from this vector.

```
##  [1]  4 -1  2 -4 -3  2 -3 -2 -4  0  0  5  1  5  2  4  0  0  4  4
```

```
## Positive numbers in this vector are:  4 2 2 5 1 5 2 4 4 4
```

Question 3: Perform the same task without the loop using logical indexing instead.

```
##  [1] 4 2 2 5 1 5 2 4 4 4
```

```
## Positive numbers in this vector are:  4 2 2 5 1 5 2 4 4 4
```

Question 4: Write a function that tests if the vectors have negative elements, and prints an appropriate message.Test the negativity of these three vectors to show the function works correctly.

```
## [1] "The vector does not have any negative elements"
```

```
## [1] "The vector does not have any negative elements"
```

```
## [1] "The vector has negative elements"
```

Data Exploration

Basic Data Description

Question 1: Before even looking at the data, what do you think, were you be able to answer the question if you have access to a suitable dataset and respective analysis tools? What might the answer look like? Maybe you know what the answer is?

Before even looking at the data, I am not able to answer the above question. However, if provided with suitable data set and respective analysis tools, I would be in a better position to derive an answer. But again, the answer to the question would vary according to the definition of dangerous. Is it more dangerous if the fatality of the attacks are more? Or is it more dangerous if the number of attacks are more irrrespective of the fatality?

Question 2: Load Data and find out the number of variables and cases

```
## The number of variables are:  24
```

```
## The number of cases are:  25841
```

Question 3: Look at the variable names. Do you understand what do they mean? Which variables do you think we need to answer the question, stated above? Do you think we have sufficient amount of data? Anything else you notice here?

```
##  [1] "Case Number...1"        "Date"                   "Year"
##  [4] "Type"                   "Country"                "Area"
##  [7] "Location"               "Activity"               "Name"
## [10] "Sex"                    "Age"                    "Injury"
## [13] "Fatal (Y/N)"            "Time"                   "Species"
## [16] "Investigator or Source" "pdf"                    "href formula"
## [19] "href"                   "Case Number...20"       "Case Number...21"
## [22] "original order"         "...23"                  "...24"
```

Yes, the variable names are very starightforward and I am able to understand the data and meaning associated with them. From my understanding, the variables needed to answer this questioon are Type, Country, Injury,and Fatal (Y/N). In my opinion, we have sufficient data to deterimine which of the two countries is more dangerous in terms of shark attacks.

Explore the data

Question 1: How many different countries are listed here in data?

```
## Number of different countries listed in the data are: 212
```

Question 2: Browse the country names. Comment on what do you see. Do all the names make sense?

While browsing through the country names, I notice some inaccuracies and incorrectness of data which do not make sense. There are continents like “Asia”, oceans like “Indian ocean”, unclear locations like “Between Portugal & India” and “Italy / Croatia, and special characters like”Red Sea?” and data like “EQUATORIAL GUINEA / CAMEROON” and “NA” which do not provide any insights.

Quetsion 3: Next, let’s look at the year of the attack (variable Year). What is it’s data type? Does it correspond to what you expect?

```
## The data type of the 'Year' column is: character
```

The datatype of the Year variable is character. Although not incorrect, it is not what I expected. I expected it to be an integer.

Question 4: How many missing values for Year do we have in data? What does this suggest about the observations and data quality?

```
## The number of missing values for the Year variable is: 19038
```

This tells us that more than 70% of the data is missing the year value and the data quality is low.

Question 5: Find the minimum, maximum, and median value of Year.

```
## The minimum value of Year is: 0000

## The maximum value of Year is: 2022

## The median value of Year is: 1983
```

Question 6: The minimum value "0" looks like a different code for missing data... So let's take a closer look. Browser the value of Date for cases where Year = 0. Comment what do you see. How many such cases do we have? what does this tell about the scope of this dataset?

For cases where Year=0, the dates are not providing a definite time/year. Words like 'before' and 'after', 'during' are used to define point of time. For example "Before 1958', 'No date',"World war'. Since there is no definite year provide, it provides ambiguity in the data and reduces the scope of the dataset.

```
## The number of such cases in the data file are: 129
```

Question 7; One of the oldest dates there is "Ca. 725 B.C.". Explain what happened and what is the source of information. What does this suggest about the used data sources?

```
## [1] "Extraction of the row having date as Ca. 725 B.C"

## # A tibble: 1 x 24
##   Case Numbe~1 Date  Year  Type  Country Area  Locat~2 Activ~3 Name  Sex   Age
##   <chr>        <chr> <chr> <chr> <chr>   <chr> <chr>   <chr>   <chr> <chr> <chr>
## 1 725/BC       Ca. ~ 0000  Sea ~ ITALY   Tyrr~ Krater~ Shipwr~ males M     <NA>
## # ... with 13 more variables: Injury <chr>, `Fatal (Y/N)` <chr>, Time <chr>,
## #   Species <chr>, `Investigator or Source` <chr>, pdf <chr>,
## #   `href formula` <chr>, href <chr>, `Case Number...20` <chr>,
## #   `Case Number...21` <chr>, `original order` <dbl>, ...23 <lgl>, ...24 <lgl>,
## #   and abbreviated variable names 1: `Case Number...1`, 2: Location,
## #   3: Activity
```

It was a sea disastor where shipwrecked sailors were attacked by sharks. The data sources used was V.M. Coppleson (1958), p.262, et al

Clean Data

Question 1: Now let's look at whether the attack was fatal (the variable Fatal Y/N. As the first step, rename this variable to something more suitable, e.g. fatal. We ask you also to only keep variables you need below and drop all the others. (You may want to return to this question later and add/remove additional variables.) You may also rename other variables if you wish

```
##  [1] "Case Number...1"        "Date"                   "Year"
##  [4] "Type"                   "Country"                "Area"                  
##  [7] "Location"               "Activity"               "Name"                  
## [10] "Sex"                    "Age"                    "Injury"                
## [13] "Fatal"                  "Time"                   "Species"               
## [16] "Investigator or Source" "pdf"                    "href formula"          
## [19] "href"                   "Case Number...20"       "Case Number...21"      
## [22] "original order"         "...23"                  "...24"
```

```
## [1] "Type"    "Country" "Fatal"   "Year"
```

Question 2: Lets focus on reasonable recent time span only. Only keep the reasonably recent cases based on the year variable. Explain your reasoning when selecting the time span.How many cases are you left with?

```
## 
## 0000 0005 1788 1791 1803 1804 1807 1825 1831 1832 1836 1837 1839 1840 1841 1842 
##   26    1    1    1    1    1    1    2    1    2    1    1    2    2    1    1 
## 1845 1847 1849 1852 1853 1855 1856 1858 1860 1861 1862 1863 1864 1865 1866 1867 
##    1    3    4    4    2    4    1    1    1    1    3    4    1    2    1    1 
## 1868 1869 1870 1871 1872 1873 1874 1875 1876 1877 1878 1879 1880 1881 1882 1883 
##    2    1    3    4    1    2    2    4    4    9    3    2    4    1    6    5 
## 1884 1885 1886 1887 1888 1889 1890 1891 1892 1893 1894 1895 1896 1897 1898 1899 
##    3    3    8    4    6    2    4    1    4    2    2    7    4    1    2    4 
## 1900 1901 1902 1903 1904 1905 1906 1907 1908 1909 1910 1911 1912 1913 1914 1915 
##    6    3    5    4    1    8    6    5    2    6    4    5    6    4    6    6 
## 1916 1917 1918 1919 1920 1921 1922 1923 1924 1925 1926 1927 1928 1929 1930 1931 
##    8    1    3    7    9    6   10   14    8    4    7   13   12   24   13    9 
## 1932 1933 1934 1935 1936 1937 1938 1939 1940 1941 1942 1943 1944 1945 1946 1947 
##   11   12   18   19   22   18   11    9   17    4   13    5   12    2   19   20 
## 1948 1949 1950 1951 1952 1953 1954 1955 1956 1957 1958 1959 1960 1961 1962 1963 
##    8   15   18   14    6   14    7   10   13   13   13   25   30   34   31   22 
## 1964 1965 1966 1967 1968 1969 1970 1971 1972 1973 1974 1975 1976 1977 1978 1979 
##   19    9   21   15   13    9    2   11   14    9   16   24   17   16   11    8 
## 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 
##   12   12   11   18   15   11   18   12   16   22   15    9    9    9   12   11 
## 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 
##   16   14   24    9   23   16   18   17   27   30   26   22   28   40   25   32 
## 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 
##   31   28   35   41   36   30   44   27   31   28   23
```

As we can see from the above table, the number of incidents taking place in Australia and South Africa regions increased from 2004. Thus, I will be taking 2004-2022 time span from here on.

```
## # A tibble: 2,188 x 4
##    Type       Country      Fatal Year 
##    <chr>      <chr>        <chr> <chr>
##  1 Provoked   USA          N     2022 
##  2 Unprovoked AUSTRALIA    N     2022 
##  3 Unprovoked AUSTRALIA    N     2022 
##  4 Unprovoked USA          N     2022 
##  5 Unprovoked SOUTH AFRICA Y     2022 
##  6 Unprovoked BAHAMAS      Y     2022 
##  7 Unprovoked USA          N     2022 
##  8 Unprovoked AUSTRALIA    N     2022
##  9 Unprovoked AUSTRALIA    N     2022
## 10 Unprovoked USA          N     2022
## # ... with 2,178 more rows
```

```
## The number of cases that I am left with are: 2188
```

Question 3: What kind of different values do you see in the fatal variable? Comment the values you see. Do you have an idea why do you see some of these figures?

```
## Different values in the Fatal variable are: N Y M n NA Nq F UNKNOWN 2017
```

Question 4: Now let's convert the fatal column into a logical variable: TRUE if the attack was fatal and FALSE if not. Convert the cases where you are unsure into missings. Explain your for decisions you make here.

```
## # A tibble: 2,188 x 4
##    Type       Country      Fatal Year 
##    <chr>      <chr>        <chr> <chr>
##  1 Provoked   USA          False 2022 
##  2 Unprovoked AUSTRALIA    False 2022 
##  3 Unprovoked AUSTRALIA    False 2022 
##  4 Unprovoked USA          False 2022 
##  5 Unprovoked SOUTH AFRICA True  2022 
##  6 Unprovoked BAHAMAS      True  2022 
##  7 Unprovoked USA          False 2022 
##  8 Unprovoked AUSTRALIA    False 2022 
##  9 Unprovoked AUSTRALIA    False 2022 
## 10 Unprovoked USA          False 2022 
## # ... with 2,178 more rows
```

I have converted all the missing values to NA. The reason being that I don't have enough clarity on the impact they will have on the data quality if they are changed to either 'TRUE' or 'FALSE'. Doing so coulld give rise to inaccuracies thus, I think it is better if these data are not considered at all.

Austalia or South Africa?

Question #1: Filter the data to only contain cases from these two countries. How many cases do you have from each country? Which percentage of those is fatal?

```
## 
##    AUSTRALIA SOUTH AFRICA 
##          453          131
```

```
## [1] "Overall Fatality Percentage of Australia:"
```

```
## 
##     False      True 
## 0.8791822 0.1208178
```

```
## [1] "Fatality Percentage of Australia: "
```

```
## 
## False  True 
##   0.9   0.1
```

```
## [1] "Fatality Percentage of South Africa:"
```

```
##
##     False      True
## 0.8050847 0.1949153
```

Question 2: Now try to answer the question: which country is more dangerous? Are you able to answer it? What do you think, what can this analysis and your answer be used for? Explain your reasoning.

The percentage of fatalities is greater for Australia as compared to South Africa. Thus, according to me Australia is more dangerous than South America. My analysis can be used for digging further into the detais and incorporating the 'TYPE' in the analysis to improve it because in my opinion, as provoked attacks should not be accounted when comparing the hazardousness.

Question 3: Finally, returning to your analysis and the original data (not the one you have cleaned), do you see any ethical issues here? Can your results be misused? Can this data used in a harmful way?

I do see ethical issues in this data since there are personal detail's of incidant loggers such as their name, gender, and etc. These data are private information and can be used for the wrong reasons in various ways. I believe, my conclusion is based on a surface level analysis. A deeper analysis may provide a different answer. Thus, my results could be misused to spread information which may not be entirely true.