EIGHTY20
T: 021 461 8020 | E: info@eighty20.co.za

THE EIGHTY20 DATA SCIENCE CHALLENGE

INSTRUCTIONS

Please complete the following set of challenges. Our preference is that you work in python in a separate jupyter notebook for each question, but you're welcome to use any other language or framework e.g. R in R Markdown files, as long as:

1. We're able to reproduce your workings by running whatever code you submit
2. It's clear how to setup the environment including any packages or libraries required
3. You communicate the what, how and why of what you're doing in each problem

We're really interesting in knowing not just that you can use code and statistical knowledge to solve data, analytics and business problems, but also how you go about solving those problems. With that in mind, err on the side of over-communicating what you're doing rather than under-communicating. You can assume your audience has the knowledge of a 2nd year Statistics and Computer Science student (understands foundational concepts and definitions but not anything very advanced).

UNDERSTANDING THE DATA

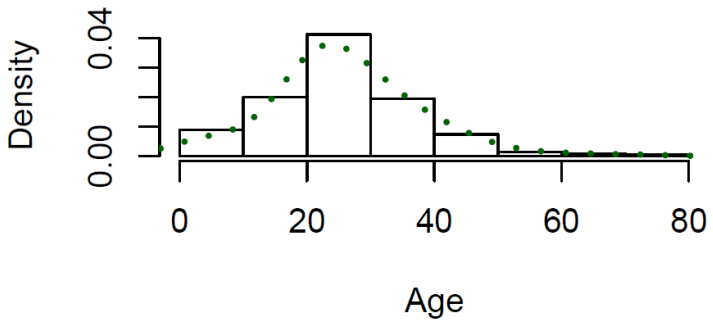
One of the first datasets that any analyst in the Big Data sphere gets introduced to is the Titanic dataset. It also regularly features in computational prediction competitions as seen on https://www.kaggle.com/.

The sinking of the RMS Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage from the UK to US, the Titanic sank after colliding with an iceberg. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class.

1. Load "titanic.csv", we want to see the deaths of passengers as per class in order to see whether preference was given to different classes. One of the best ways to visualize this is by simply plotting the data. Construct a plot to display deaths and survival per class.
2. Next, aggregate the data to see the gender split that was on the boat. What interesting feature do you notice when getting the gender split by class?

Now that you have a better idea of the data structure, lets get more familiar with the statistical properties of it. One of the columns contain the age of the passengers

1. Plot the histogram of the age per passenger class
2. Plot the histogram with its appropriate density over histogram
3. The figure below represents the age distribution of the male passengers in third. Why do you think this right-tailed distribution exists for this specific demographic segment? What story does this tell?



Hearing of the tragedy, the first comment you often hear when the discussion of the life boats come up is that women and children were first.

1. Does this reflect in the data? Make the assumption that a child is a passenger under the age of 15.

PREDICTIVE ANALYTICS

A lot of what we do here at Eighty20 employs using prediction techniques. Using the titanic dataset, use any statistical model or method to predict who was most likely to survive. Show clearly how you evaluate how well your model has performed and comment on any potential problems with your approach.

USING DATA TO ADDRESS A QUESTION

The hflights dataset can be considered to be a large dataset consisting out of 227 496 rows and 21 columns. This dataset contains all flights departing from Houston airports IAH (George Bush Intercontinental) and HOU (Houston Hobby). The dataset can be obtained from the

following URL. Use your preferred method to grab the data but clearly show how you do so in your code

https://github.com/selva86/datasets/blob/master/hflights.csv

1. How many unique carriers exists in the dataset?
2. Using the average departure delay, create a boxplot to visualize the distribution of delay per airline Carrier (limit the sample to delays smaller than 60min). Which airline was the best to fly with and which one was the worst? What was the median delay for both?

APPROACH TO A GENERIC PROBLEM

By starting at the top of the triangle below and moving to adjacent numbers on the row below, the maximum total from top to bottom is 23

```
          6
        7   8
      9   2   8
    9   4   9   4
  8   9   8   6   0
```

```
            40
          34    33
        27    19    25
      18    13    17    10
    8     9     8     6     0
```

That is, 3 + 7 + 4 + 9 = 23.

Now find it for triangle:

```
                75
              95 64
            17 47 82
          18 35 87 10
        20 04 82 47 65
      19 01 23 75 03 34
    88 02 77 73 07 63 67
  99 65 04 28 06 16 70 92
41 41 26 56 83 40 80 70 33
41 48 72 33 47 32 37 16 94 29
53 71 44 65 25 43 91 52 97 51 14
70 11 33 28 77 73 17 78 39 68 17 57
91 71 52 38 17 14 91 43 58 50 27 29 48
63 66 04 68 89 53 67 30 73 16 69 87 40 31
04 62 98 27 23 09 70 98 73 93 38 53 60 04 23
```

As there are only 16384 routes, it is possible to solve this problem by trying every route and finding out which route yields a maximum of all routes (Note: Picking the largest adjacent number will not necessary lead to a maximum outcome for the path in total).

Find the maximum total from top to bottom in "triangle.csv"