[redacted]
ANALYTICS

Data Science Task Sheet

Introduction

On the following pages are two exercises for you to address to give you a flavour of topics you'll encounter as one of our data scientists. Your answers will also show us something of the way you work, and perhaps, how you think. Almost all questions do not have a definitive right answer and there are no restrictions to non-human resources you can use. The questions are designed to be self-explanatory and straightforward, but if you're not sure what is being asked, you can contact Colin: [redacted]
We do require that it is all your own work.

In addressing the exercises, you can use whichever software you like. We'd like you to present concise, efficient, easy-to-read, annotated code wherever you use it, and allow us to reproduce your results. Where you describe your findings, do so in a way that a non-expert could easily understand, being careful with wording.

We expect it to take perhaps half a day. You should submit answers to questions and your code; not relying on us having to run your code to see your answers. It's fine to submit a single document (e.g., .ipynb or html/pdf you've created from .Rmd) or separate code/answer documents as you prefer. We don't want any data files, or individual images. If your submission is larger than a couple of megabytes, you're probably including more than you need to. If you submit several files, it is best to put them in a folder and compress it before sending, but please avoid over-long folder/file names as these can sometimes cause extracting to fail. Please name your submitted (+/− compressed) file with your name and up to five letters as a suffix which indicate which recruitment agency sent you the task sheet, e.g., 'Joe Smith JOBZZ.zip'. Please don't send links to external storage as our security systems often don't like those.

1 k - anonymity analysis

For this exercise, you are advised to read all parts first since the example in Part d) will help you understand what is required for Part a).

a)

Randomly generate a dataset (dataframe) with eight columns and 50,000 rows. Each column should be a categorical variable (of arbitrary name) with three levels (of arbitrary names) in approximately equal proportions.

b)

Verify that the proportions of each value are similar for each of the eight columns.

c)

How many unique rows (i.e., permutations of category levels) are possible?

d)

Produce a table and appropriate graph which show the frequencies (numbers of groups) by permutation group sizes up to group size of 12. That is, how many groups are unique combinations (group size = 1), how many groups are made up of a pair of matching combinations (group size = 2), how many groups are made up three the same, etc?

For example, in Table 1 (conveniently ordered) of three columns and eight rows, there is one unique row (one group of 1), four rows in pairs (two groups of 2) and three rows in groups of three (one group of 3). Each column has three possible values, which just so happen to be the same (*a, b* and *c*), but they don't need to be. Table 2 shows the corresponding frequency table, which summarizes the group sizes; it is this that you are asked to produce for the data you created in part a).

e)

Comment upon the distribution of group sizes in d).

Table 1 – An example table of three categorical variables

X	Y	Z
a	a	b
a	a	b
a	a	b
b	c	c
b	c	c
c	a	c
c	a	c
c	b	a

Table 2 – The corresponding frequency table which summarizes group sizes of Table 1

Group Size	No of groups
1	1
2	2
3	1

f)

If your random variables were, in fact, meaningful information on individuals, which group sizes are of most concern from a privacy perspective?

g)

Consider the effect of missing data in the dataset you created in Part a). How might this complicate the production of a frequency table of group sizes in Part d)?

h)

Imagine the code that you wrote for Part d) was to be deployed in an automated system that [redacted] customers could use independently, on potentially large volumes of their own data. Describe how you might deploy the code, and what additional considerations you might have or any changes to the code you might make. Note: it is not necessary to provide another version of the code created for d).

2 Postcodes and Privacy

In the US, 5-digit Zip codes are usually rounded to 3-digits when anonymizing health data, so knowledge of the Zip code doesn't allow small groups to be identified. Even then, there are some 3-digit codes that have fewer than 20,000 residents, and the advice is to lump these together under a new code (000).

Looking forward to how GDPR may affect data handling in the UK, might a similar approach be possible here? In answering this, use the data below. You might want to include some examples of any postcodes which could be problematic. Write it up as a mini report to inform the decision of a privacy officer.

UK population by postcode data (28 MB) found here:
`www.nomisweb.co.uk/output/census/2011/Postcode_Estimates_Table_1.csv`.