Mirador Analytics

Data Science Task Sheet

Introduction

Below are two exercises for you to address which will 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. There are no definitive right answers (except for 1c), there are no restrictions to resources you can use and the questions are not designed to 'catch you out', so if you're not sure what is being asked, feel free to contact Colin: colinmoffatt@miradoranalytics.com. 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 easy-to-read, annotated code wherever you use it, allowing us to reproduce your results. Where you present your findings, do so in a way that a non-expert could easily understand.

It should take a half-day at most, and we'd expect no more than four pages.

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 roughly 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)

Write some code to produce a table and/or graph which shows the frequencies (numbers of rows) by permutation group sizes (up to group size of 10). That is, how many rows are unique combinations (group size = 1), how many rows are one of a pair of matching combinations (group size = 2), how many rows are one of a group size of three, etc?

For example, in the table below left (conveniently ordered in groups) of three columns and eight rows, there is one unique row, four rows in pairs and three rows in groups of three (just one group, in fact). Each of the variables in the table has three levels; a, b and c. The table on the right shows the corresponding frequency table, which you should produce for the data you created in part a).

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

Group Size	No of rows
1	1
2	4
3	3

e)

Comment upon the distribution produced in d).

f)

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

2 Postcodes

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, and write it up as a mini report. UK population by postcode data (28 MB) found here:

`www.nomisweb.co.uk/output/census/2011/Postcode_Estimates_Table_1.csv`.