PrivacyHub by DATAVANT

Data Science (Privacy) Task Sheet

Introduction

On the following pages is an exercise for you to address to give you a flavour of topics which will concern you as one of our privacy data scientists. Your answers will also show us something of the way you work and think as well as how well you communicate. The questions are designed to be self-explanatory and straightforward, but if you're not sure what is being asked, you can contact Colin: colinm.privacyhub@datavant.com. We do require that it is all your own work.

In addressing the exercises, you should use R or Python. Where you describe your findings, do so in a non-technical way, being careful with wording.

You should submit answers to questions and your code; not relying on us having to run your code to see your answers. Do not reproduce the question. 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 through which avenue you are applying, e.g., a recruitment agency or a web site, e.g., 'Joe Smith JOBZZ.zip'. Please don't send links to external storage as our security systems often don't like those.

US ZIP Code Populations

The following questions involve the downloading and joining of two datasets. This is open source data available on the internet which is a few years out of date. Read through the whole of the question first so you can remove any columns from the data that are not required in advance. The dataset at tinyurl.com/TaskDataA[1] has a number of variables including population by 5-digit ZIP code for (nearly) the whole of the USA. Zip code, or more properly ZCTA (Zip code tabulation area) is the ninth column. This dataset, however, does not have the US state of each ZIP code (despite the description implying it might), but the one at tinyurl.com/TaskDataZ does (ZCTA is the second column), so download and (inner) join the datasets.

For the following questions, where appropriate, use code that is concise, efficient, easy-to-read and annotated to explain what's going on. Where the answer is a 5-digit ZIP code, also give the state.

a)

Which state has the most 5-digit ZIP codes, and how many such codes does it have?

b)

Give the 5-digit ZIP code which is i) the most easterly, and then excluding Alaska ii) the most westerly, and iii) the most northerly.

c)

Which 5-digit ZIP code has the highest population density? Present that density in a sensible, informative way.

d)

Below you will create another dataset where you aggregate 5-digit ZIP codes to 3-digit ZIP codes (based upon common first three digits), but before you do, you need to check state and ZIP alignment:

[1] A fairly poor data dictionary can be downloaded here: https://tinyurl.com/TaskDataADesc

i) Which 3-digit ZIP codes are common to more than one state?

ii) Name the incongruous places (city)'. Some health datasets have state and 3-digit ZIP code of patients alongside sensitive health information (and other demographic information such as sex, age and race). Explain why the residents of these particular places have a relatively high risk of re-identification in such datasets.

iii) What modification of such datasets would you suggest to reduce this risk and why?

e)

Create a new dataset based upon what you already have, which contains columns showing 3-digit ZIP codes, state, population and land area (exclude the places you found in d).

i) How many 3-digit ZIP codes have a population smaller than 20,000 residents (ignore zero populations and assume any with single digits are also zero).

ii) Standard guidance to reduce risk in such 'small' 3-digit ZIP codes is to amalgamate them into a single group coded as '000'. What are the pros and cons of this approach?

iii) Can you suggest a superior approach?

iv) Produce a sensible plot to visualize the variation in population density of the 'small' 3-digit ZIP codes and comment upon it.

v) Where is the 3-digit ZIP code with the smallest population density? Is this a surprise?

f)

Write a short illustrated account which compares the distributions of population sizes for 5-digit and 3-digit ZIP codes, commenting upon any patterns or relationships. Present any code you use and analyses as an appendix to your 'report'.