MATH 70096 - Exploratory Data Analysis & Visualisation

Spring 2022 - Assessed Coursework 4

Deadline: 15 March 2022, 23:59 (UK time)

You should submit an R Markdown notebook, containing your answers to these questions, via the Imperial College Blackboard VLE, by the deadline stated above. Your submission should include both the .Rmd source file for your notebook, as well as the corresponding, compiled PDF document.

This coursework should involve no more than **3 hours** of effort. The available marks are indicated in square brackets for each part of the question.

This coursework counts for 10% of your total mark for EDAV.

Plagiarism: Your submission should be *your own work*. Note that software tools are used for plagiarism detection.

Q1) Consider the data contained within the dataframe `boston.c`, which is provided as part of the spData package in R. This dataset contains housing data that was collected as part of the 1970 census of Boston, Massachusetts. Each observation (row) in the dataset contains a collection of statistics corresponding to a single census 'tract' (a small geographic region containing multiple houses, defined specifically for a census).

This question will consider the spatial distribution of the `CMEDV` variable. This variable corresponds to the median value (in USD 000s) of owner-occupied housing in each census tract. Each tract is also associated with a point location; geographic coordinates for this point (measured in decimal degrees latitude and longitude), as well as the town in which it is located (within the Greater Boston area), are provided for each observation.

To start, load the dataframe in `boston.c` and use dplyr verbs to create a smaller dataframe, `BostonData`, which contains only the variables `TOWN`, `LON`, `LAT` and `CMEDV`.

a) Construct a visualisation that illustrates the spatial distribution of the observations in `BostonData`. Provide a brief statement explaining why your visualisation demonstrates that the coordinates that are given in the dataset must be incorrect.

[2]

It has been suggested that the coordinates provided in `boston.c` contain a systematic error, with all observations in each town being mislocated by a fixed distance (in a fixed direction).

Suppose there are n_j observations in town j, and for observation k in town j, denote the longitudinal coordinate $x_{j,k}$, $k = 1, ..., n_j$. Then we assume

$$x_{j,k} = TC_j^{(x)} + \Delta_{j,k}^{(x)},$$

where $TC_j^{(x)}$ is the longitudinal coordinate of the centre of town j, and $\Delta_{j,k}^{(x)}$ is the displacement of observation k in town j from the town centre. Suppose the latitudinal coordinates (which we denote $y_{j,k}$) satisfy a similar relationship. The suggested systematic error is therefore such that $\left(TC_j^{(x)}, TC_j^{(y)}\right)$ has been misspecified for $j = 1, ..., n$, where n is the number of towns.

The accompanying file, `BostonTownCentres.csv` contains the correct coordinates for each town centre in Boston.

b) Load the town centre coordinates into a dataframe and use an appropriate mutating join to combine this with `BostonData`. Use a similar visualisation as in part (a) to illustrate the spatial distribution of town centres. **[3]**

c) For each town (i.e. for $j = 1, ..., n$), replace the centroid of the n_j `boston.c` locations with the true town centre. Hence add two new columns to your combined dataframe, containing the true coordinates for each observation. **[2]**

d) Hence construct a visualisation that shows the spatial distribution of the median value of owner-occupied housing in Greater Boston in 1970. **[1]**

Q2) Using an appropriate dataset from the spData package in R, construct a choropleth map that visualises the state-level median income across the United States, as reported in 2010. **[2]**