MA20277 2022 - Coursework 2

PLEASE INSERT YOUR NAME OR CANDIDATE NUMBER HERE

Question 1 [9 marks]

We want to analyze the books “Anne of Green Gables” and “Blue Castle” by Lucy Maud Montgomery. The two books are provided in the files “Anne of Green Gables.txt” and “Blue Castle.txt”.

a) *Visualize the frequency of the 10 most frequent words that satisfy the following three criteria: (1) The word occurs at least five times in each book, (2) The word is not a stop word according to the usual stop list considered in the lectures, (3) The word is not “I’m”, “don’t”, “it’s”, “didn’t”, “I’ve” or “I’ll”.* [**6 marks**]

b) *Some scholars say that “Anne of Green Gables” is patterned after the book “Rebecca of Sunnybrook Farm” by Kate Douglas Wiggin. The text for “Rebecca of Sunnybrook Farm” is provided in the file “Rebecca of Sunnybrook Farm.txt”. Extract the top two words with the highest term frequency-inverse document frequency for each of the two books, “Anne of Green Gables” and “Rebecca of Sunnybrook Farm”, with the corpus only containing these books.* [**3 marks**]

Question 2 [9 marks]

We were given PM10 measurements from 60 measurement stations in the Greater Manchester area, including the locations of the stations. The data can be found in the file “Manchester.csv”. A detailed description of the variables is provided in the file “DataDescriptions.pdf”.

a) *Visualize the data in an informative way and provide an interpretation of your data graphic.* [**3 marks**]

b) *Explore the spatial dependence of the PM10 measurements.* [**3 marks**]

c) *Provide estimates of PM10 levels for two locations: (1) Latitude=53.354, Longitude=-2.275 and (2) Latitude=53.471, Longitude=-2.250. Comment on the reliability of your estimates.* [**3 marks**]

Question 3 [28 marks]

After hearing about the work you did for Utopia’s health department, the country’s police department got in touch. They need help with analyzing their 2015-2021 data regarding certain crimes. The data is provided in the file “UtopiaCrimes.csv” and a detailed explanation of the variables is provided in the file “Data Descriptions.pdf”.

Utopia consists of 59 districts and a shapefile of Utopia is provided together with the other files. To hide Utopia’s location, the latitude and longitude coordinates have been manipulated, but the provided shapes are correct. The districts vary in terms of their population and the population for each district is provided in the file “UtopiaPopulation.csv”.

a) *What are the three most common crimes in Utopia? Create a map that visualizes the districts worst affected by the most common crime in terms of number of incidents per 1,000 population.* [**5 marks**]

b) *You are told that District 44 is notorious for drug possession. The police is planning to conduct a raid to tackle the issue, but they are unsure on which area of the district they should focus on. Help them make the correct decision.* [**5 marks**]

c) *The police would also like to understand which group of people is most at risk of a burglary. The possible victims are: “young single”, “young couple”, “middle-aged single”, “middle-aged couple”, “elderly single”* *and "elderly couple". Use the short description provided in "UtopiaCrimes.csv" to extract which group of people is suffering from the highest number of burglaries. What is the proportion of burglaries that involved more than two criminals?* [**4 marks**]

d) *Make up your own question and answer it. Your question should consider 1-2 aspects different to that in parts 3a)-3c). Originality will be rewarded.* [**7 marks**]

e) *Write a short (two paragraphs) report about the findings of your analysis in parts a-d. The report should be readable for people without data science knowledge. Make it sound interesting and state possible recommendations that may be of interest to Utopia's police department.* [**7 marks**]