CS685: Data Mining

Assignment (100 marks)

Due on: 10th November, 2024, 11:30pm

This assignment will explore the tourism data for India from the website `tourism.gov.in`. It contains data for 9 years for both incoming and outgoing tourists at a month-level and country-level granularity. In addition, it has several gender-wise, age-wise, purpose-wise breakups as well as state-wise visits.

If, for any particular year, the corresponding data is missed, you may indicate that in your submission.

Submit all the necessary components of your data as a single `zip` file named `rollno-assgn.zip` in the `canvas.cse.iitk.ac.in` portal within the deadline.

If you do not follow the naming conventions, marks for that question will be automatically 0 (zero).

The programs/scripts/notebooks should run in the Linux operating system.

1. (10 marks) For every year, find the change in incoming tourists in absolute and percentage terms from the last year.

   Do this for every month and the year as a whole.

   Which years show the highest positive and negative changes?

   For each month, find the changes across the entire data from the previous month.

   Which months show the highest positive and negative changes?

2. (5 marks) For every year, find the lean and peak months of visit in terms of fees received.

   For every country and continent, find the lean and peak quarters of visit.

3. (15 marks) For every year, model the data as a multinomial probability distribution for the following features: (1) gender, (2) age, (3) purpose of visit. Do it overall as well as country-wise and continent-wise.

   Find the years when the entropy of the distributions (three seaparate) for the overall count is the maximum and the minimum.

   Repeat it for every country and continent.

4. (5 marks) Repeat Q1, but for outgoing tourists.

5. (5 marks) Repeat Q2, but for outgoing tourists.

6. (10 marks) Repeat Q3, but for outgoing tourists. The features for this question are (1) departure mode, (2) departure port, and (3) state-wise visit of domestic versus foreign tourists.

7. (10 marks) Find the balance of tourists per year per country and continent. Balance is the difference of number of incoming versus outgoing tourists.

8. (5 marks) Using the answer in Q7, find the top-5 countries with highest positive and highest negative balances per year.

   Order the continents in terms of this balance per year.

9. (15 marks) Impute (or interpolate) data for the following years: 2018, 2020, 2021. When you are imputing, assume that data for all the other 9 years are available, except this.

   Use at least 3 different imputation techniques.

   Impute for the following features: (1) Purpose of visit for each continent, (2) Total number of incoming tourists per continent, (3) Number of domestic toursists per state, and measure the accuracy for the 3 different techniques.

   Explain the differences between the techniques. You may impute and show the results for some more features, if you wish.

10. (15 marks) Explain the effect of Covid on the tourism data. You may assume 2020 and 2021 to be the Covid years. In addition, you may do more fine-grained analysis.

11. (5 marks) Write a manual that describes how to use your code. Include all the programs, their plugins, and dependencies needed to run the program. Include a top-level script `assgn.sh` that runs the entire assignment.

Important clarifications regarding the assignment:

1. Assume that 2015 data is absent (even if some 2014 files specify it), and ignore it in all questions.
2. The first part of second question has changed.
3. Consider China and Taiwan as part of East Asia and do calculations accordingly.
4. For 2021, consider the new port-wise distribution file, as in the new updated zip file (also posted separately).
5. For slight inconsistencies refer to the `tourism.gov.in` website and specify them in your assignment. If you change any Excel file, then submit the updated Excel file along with the solution.
6. Ignore the inconsistencies in the number of FTAs because some tourists are not classified under any category.
7. For the purpose of visits varying in different years, rename similar purposes put them under the same umbrella, and specify it in your solution
8. Naming errors like the UK and the United Kingdom need to be handled by you
9. In each of 2013, 2014 and 2018, one table is missing; so, for these years, respective questions can be skipped.
10. If there is any question that cannot be completed due to lack of data for some year, please mention it clearly in the solution.