Assignment-1

Data Mining
Report

Submitted By:
Nitesh Malviya
(150465)

Folders containing:

- Datagov:data given on zip
    - Demography: contained Python script to clean and merge data of this category
    - Economy: same as above
    - Education: same as above
- Q1: containing data and code for question 1 (same goes for all other Qx)

Folder:

- Datagov -> Demography
    - ch.py - All the csv files in this folder are stored as string, then the states column is changed so that it can be merged with regions.csv file properly. when the files are merge, missing NaN of particular region is found and replaced with the mean of particular region and giving a single merge file for all 3 files.
- Datagov -> Economy
    - economy.py - the states of all the data given in this folder were in index so have to convert them in rows, loaded all the data files in the start, one row was redundant so removed it via index = 11 and added items description and year to make it one column so transpose can be done properly, then added all region manually(south, ne, etc)
    - Then classified the data by region with a loop and replaced the NaN value with the mean of respective region
    - There might be some data with still NaN as the data of whole region may be missing, that data will be consider later when all the csv files are merged to one
- Datagov -> Education
    - two.py - solved this folder considering 4 types of data, first - literacy rate as it didnt contain any year, second - drop rate as it contains 2012-15 years data , thirdly higher education which contained 2010-16 years data and lastly all the rest data which were all 2013-16 data.
    - Classified all the data into region then year basic and then taking the mean to fill the NaN values, then added the year to the Index so that all the data can be merger easily, this process is done for all the 4 types of data just taking different year
    - And finally merging all the data of this folder
- Q1 -
    - Takes all the data and merge them, it is done by combining all the codes mentioned above (two.py,ch.py,economy.py)

    - Final_merge.csv is created which contains the final merged and clean data
    - All the missing data whos whole region data was not available was filled by taking the mean of all the states and filling it at that place
- Q2 -
    - Final_merge.csv from previous file is used to do this problem
    - q2_w_norm.py - this codes takes the norm of all the vector by a built in sklearn library and finally getting norm value for all, then finding the cosine similarity by taking dot product with all states and All India data and then nearest to 1 is found out hence giving the 5 similar states which comes out to be - Tamil Nadu, NCT of Delhi, Haryana, Chhattisgarh, Odisha

- Q2 -
    - q2_wo_norm.py - euclidean distance is found out by taking the different of each row with the All india row then squaring the data adding then taking square root and the value smallest is near to the All india hence giving Maharashtra, Tamil Nadu, Uttar Pradesh, Gujarat, Karnataka
- Q3 -
    - Simply takes Correlation of the merge_final.csv
- Q4 -
    - q4_all_data.py - this takes the data of final_merge.csv and converting the numbered row and column to list of list and then converting it to np array then finding the norm column wise, now a random number is generated from 1-37, 1000 times and each time it finds the nearest hit and nearest miss by taking the random number and finding the corresponding index to the csv files of final merge and then finding the region of that random number and then storing all the index value of the same region in a list so the nearest hit can be found taking that the euclidean distance with all the other values in list other then random number and finding the index of the nearest hit and same is done with other values not in list from 1-36 hence making it other region and calculating the same and fining nearest miss
    - Same concept is used for all different files
    - After finding the most significant, those 2 column are added to a new csv files
- Q5 -
    - Taking the two most significance csv files from Q4 folder and then simply using matplotlib.pyplot to plot the scatter plot