README FILE

Title: Ranking Cities in the world based on Lifestyle

LIBRARIES NEEDED

We will need the following libraries to execute the given code:

```
from random import randint
import json
import pandas as pd
import numpy as np
from pprint import pprint
```

The file format we used to input the files in the code was in json format.

SOURCE CODE DETAILS

The country indices was collected using the **cities_list.py** using the Api provided by numbeo.com.

new.json file will be created after running the cities_list.py

DATA SET DETAILS

Each json object in new.json file represent each city with 18 attributes.

new.json file is divided in to three set of dataset based on threshold value equal to 8 , namely fullAttrU.json-which contains all cities indices with 18 attributes,lessAttribute.json-which contains all the cities inidices with >=8 and <=18 attributes and leastAttribute.json which contains all the city indices with <8 attributes .leastAttribute.json is discarded since

new.json file is divided using dataextraction.html where new.json is take as input to this html file

we have used HFS file server to run the html file.

We have used below mentioned simulator to create the correlation matrix

http://www.sthda.com/english/rsthda/correlation-matrix.php

with the help of the matrix we have created correlation.txt file manually

the correlation.txt file is used as input to correlated_list.py.

correlated_list.py is used to create 15 correlation file where each file contains the attribute with top 2 correlated value of that attribute.

regression.py is used to handle the missing data from leastAttribute.json by using correlation values

the output from regression.py is combined with fullAttrU.json to get the full data

fullAttrU.txt is used to rank the city using spark sql in java

we used eclipse as IDE to run the rank.java file which creates .csv file with city and rank

ALGORITHM DETAILS:

City indices dataset is divided in to three dataset using threshold value and missing data is handled using correlation and regression and ranked the city using Quality of life indices formula from numbeo.com

Note: need to keep all the above mentioned file in one folder

Instruction to run the code on Terminal:

1.All the python file needs to be executed Using python followed by filename.

```
Python correlated_list.py
Python cities_list.py
Python regression.py
```

2.".html" is ran using HFS file server

3. ".java" file needs to be executed using Eclipse and spark 2.4.0