ANALYSING DATASETS HADOOP & PIG

Report submission as a requirement for the module of "Advance Data Management"

MSc Degree in Data Analysis
at the school of computing
Robert Gordon University
Aberdeen, Scotland
December 2020

Introduction

The following project tested my skills in data management using Apache Pig with Hadoop in the Unix environment. I have applied analysis procedures to extract knowledge from data and efficiently handle, manipulate and implement data management techniques in large datasets.

The project is divided into two tasks. Task one required me to analyse a given dataset given by the module coordinator. I will go over each analysis and the results in further detail in the next following pages. Task two required me to choose a dataset of my liking. I have chosen a COVID-19 dataset because while this project is taking place, we are still in the middle of this pandemic and it seems relevant and an exciting topic. This task requires me to propose three insightful analysis that can provide useful information for decision making.

All the Pig scripts and results have been included as external files together with this report.

TASK 1

Question 1

Find the number of countries in the dataset..

Question 1 reults

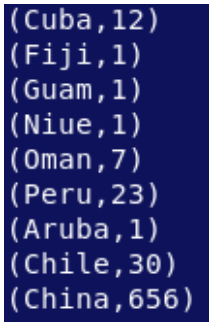
My script found to be 163 unique countries in the given dataset. I am assuming all contries have been annotated correctly.

Question 2

List the countries together with the number of cities in each country.

Question 2 results

I have included a screen capture, part of the output by my pig script. Please refer to the file ResultsQuestion2.txt in the attached files to see all results. I am assuming all countries and cities have been annotated correctly and mapped correctly.



Question 3

List countries in ascending order of female-to-male ratio, throughout the years

Question 3 results

The results below show Kuwait to have the smallest female to male ratio, followed by Qatar. The result makes sense as the smallest female to male ratio appear several times in the same cities but different years. Please refer to the file ResultsQuestion3.txt in the attached files to see all results.

```
(Kuwait,2005,0.2133228487980229)
(Qatar,2008,0.31051097890637763)
(Qatar,2010,0.31398951638664574)
(Qatar,2007,0.3446898289810908)
(Kuwait,1995,0.4481143866548903)
(Qatar,2006,0.4989132922418719)
(Holy See,2000,0.5085066162570888)
(Qatar,2004,0.511222431049877)
(Bahrain,2001,0.5601606997558991)
(Guam,2000,0.6369047619047619)
```

Question 4

List the top 10 most populated cities according to the most recent data in the dataset

Question 4 results

I am assuming each city in the dataset to be unique (i.e. There are not cities which contain the same name but belong to different countries.) If this were the case, my script would have grouped (country, city, year) instead of (city, year). According to my results "Ciudad de Mexico" in 2010 was the most populated city with 28,967,922 habitants.

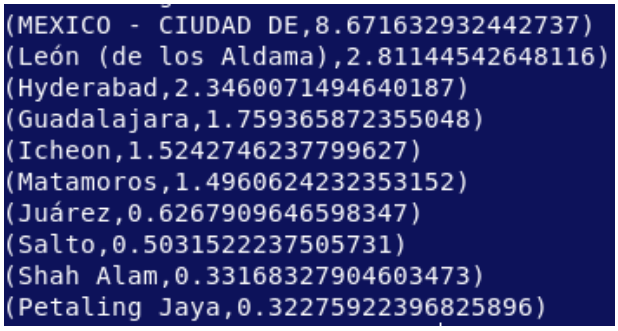
```
(MEXICO - CIUDAD DE,2010,28967922)
(Mumbai (Bombay),2001,28412836)
(Delhi,2001,22756642)
(Tlalnepantla,2010,20770252)
(Kolkata (Calcutta),2001,17778573)
(Shanghai,2000,14348535)
(Istanbul,2012,13596781)
(BOGOTA - D.C.,2005,13542016)
(PARIS,2010,12703949)
(Rio de Janeiro,2010,12640892)
```

Question 5

List the top 10 cities which have the highest population change per year in percentage since the start of the survey.

Question 5 results

As in question 4 I am assuming each city to be unique in name and belong to one country. If this is not the case then the grouping in line 30 can be changed to group country and city. The results show "Ciudad de Mexico" to have the most significant population change across the years, this result makes sense because it is also the most populated city in the world.



TASK 3

I have chosen the currently popular COVID-19 dataset found in the following source "https://github.com/datasets/covid-19" COVID-19 is an acute respiratory syndrome with a worldwide effect. The dataset lists confirm cases and reported deaths, organized by countries. I have chosen this dataset because while working on this project countries are still struggling with the virus. It seems relevant to the occasion and an interesting choice to know more about the current pandemic. Below I have included a table containing each of the attributes within the dataset and a description of each.

Attribute Name	Description
Date	Contains the cases and deaths published for a specific country in a specific date.
Day	The day the cases and deaths were published
Month	The month number the cases and deaths were published
Year	The year the cases and deaths were published
Cases	How many cases were registered on a specific date.
Deaths	How many deaths were registered on a specific date.
Country	The country where the cases and deaths occurred.
Population	The current population of that country

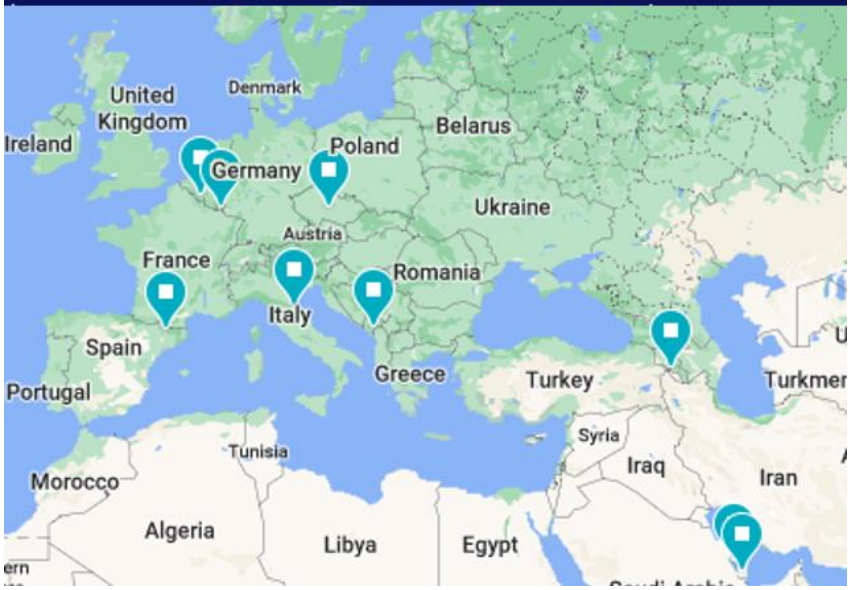
Analysis 1

Many online sources contain the countries that have been most affected and least affected by COVID 19. It is interesting to me to have a look at it from a different perspective. Seeing how every country has a diverse population density, I decided to rank the topmost and least most affected countries in proportion to their population. This analysis is interesting because most affected countries have a bigger population; hence more cases and deaths. I assume that third world countries which have inadequate health systems should be the most affected. This is why analysing in proportion to the population will allow me to have greater insight.

Analysis 1 results

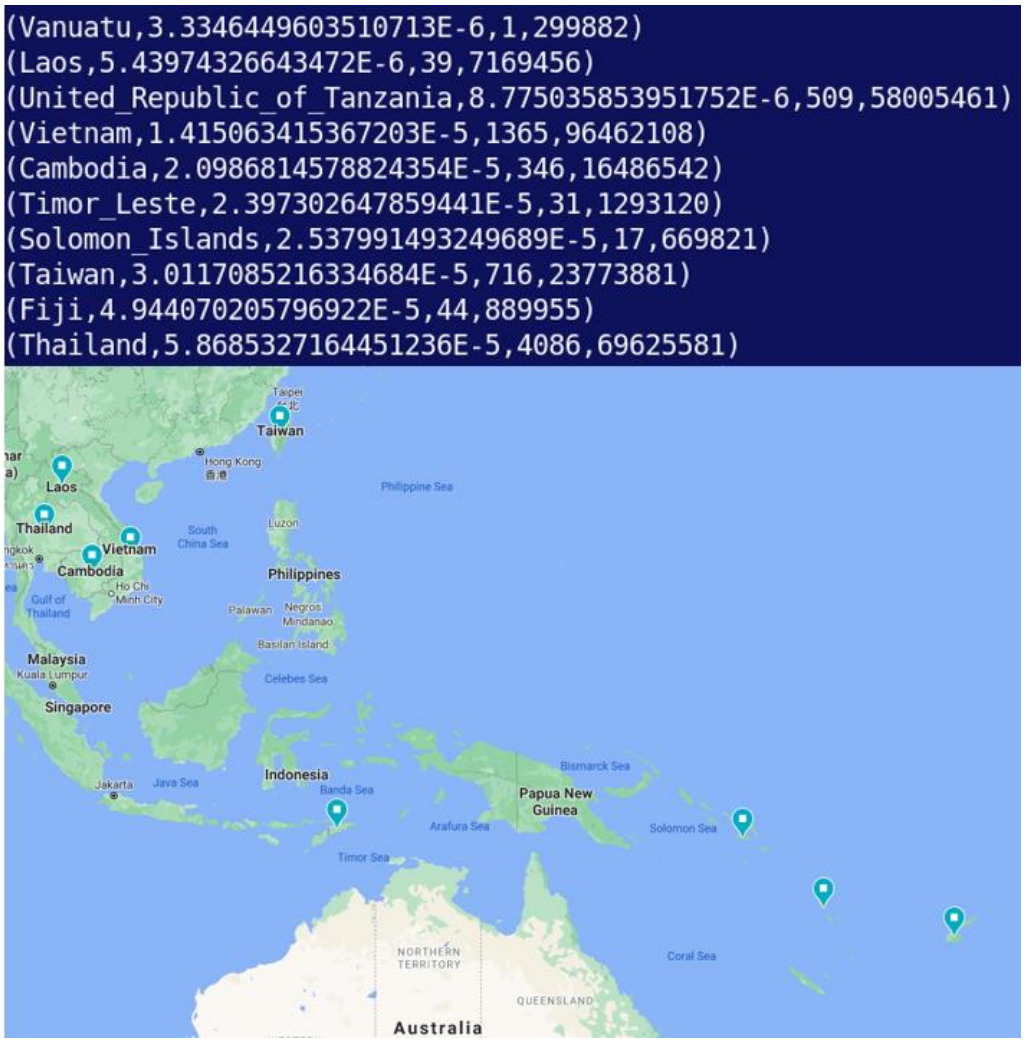
The results are interesting; I have found that most of the countries affected are landlocked. They are all very close to Italy, Spain, Belgium, France which have been severely affected. Belgium contains one of the busiest airports which can explain why the virus impacted Belgium as well as Luxembourg.

```
(Andorra,0.09195688987489663,7005,76177)
(Montenegro,0.060880899801022854,37879,622182)
(Luxembourg,0.060298683486074145,37017,613894)
(French_Polynesia,0.05409169844424155,15107,279285)
(Bahrain,0.053457180391478246,87732,1641164)
(San_Marino,0.05192581197573506,1789,34453)
(Belgium,0.051486624045580125,589806,11455519)
(Czechia,0.0510975792972638,544179,10649800)
(Qatar,0.04935716653996316,139783,2832071)
(Armenia,0.04798852362353807,141937,2957728)
```



Andorra was the worst affected, it has only 7,005 confirmed cases, but its population is 76,177 having the most cases per person. It seems most of the affected small countries were probably infected by big ones. They also seem to be in the same busy area, basically the heart of the European Union.

The second part of this analysis is also interesting. Most of the least affected countries are found around the same area as we found to be the case with the most affected countries. Half our results show small islands were not significantly affected; most of the cases are below 30; this makes sense because they are remote and not heavily transitted. Thailand, Cambodia, Laos, Vietnam all came up very positive too; they managed to handle the outbreak well even though they were close to the pandemic's epicentre; this is assuming they have run as many tests as other countries. It is comforting to see that the results returned by my analysis make sense and that all the results have in common a location.



Analysis 2

Initially, for this analysis, I wanted to find the peak month in cases and deaths for each country and have them show each on a separate tuple inside a bag belonging to the country. While working on this analysis, I realised that most peak deaths occur in the same or the next month as peak cases. I found this logical since the month when most cases happen most deaths will follow, but I noticed certain countries did not follow this pattern. I found it interesting to then direct my analysis to find the countries which had the biggest difference in the dates the country hit a peak in cases and deaths. The tuple on the left represents peak month cases, and the tuple on the right represent peak month deaths.

Analysis 2 results

The results show 85% of the countries in the data set make logical sense having the number of cases and death being the same or one month away. 15% of the countries approximately 30 of them had their cases and deaths 4 to 8 month apart from each other, it is these anomalies which seem very interesting to me. We can see South Korea reach peak deaths “147” in March while reaching peak cases “7690” in November. We found a similar pattern in the follow-up countries were the peak deaths happen during April and May, I assume due to the virus propagating quickly throughout the world during these months. We also see a pattern in peak cases, mostly in November. I assume this could be due to implemented measurements and COVID-19 tests becoming finally available worldwide by the end of the year, which allowed countries to start testing their citizens.

```
(South_Korea,{(South_Korea,11,2020,62,7690)},{(South_Korea,3,2020,147,6855)},8)
(San_Marino,{(San_Marino,11,2020,3,658)},{(San_Marino,3,2020,25,228)},8)
(Denmark,{(Denmark,11,2020,110,34127)},{(Denmark,4,2020,366,6431)},7)
(Sweden,{(Sweden,11,2020,1006,129842)},{(Sweden,4,2020,2514,16225)},7)
(France,{(France,11,2020,15760,886499)},{(France,4,2020,21063,83892)},7)
(Finland,{(Finland,11,2020,35,8719)},{(Finland,4,2020,193,3593)},7)
(United_States_of_America,{(United_States_of_America,11,2020,37165,4335894)},{(United_States_of_America,4,2020
,57796,875289)},7)
(Spain,{(Spain,11,2020,9191,462509)},{(Spain,4,2020,17203,110916)},7)
(United_Kingdom,{(United_Kingdom,11,2020,12016,627582)},{(United_Kingdom,4,2020,23999,137469)},7)
(Estonia,{(Estonia,11,2020,39,7281)},{(Estonia,4,2020,47,951)},7)
(Norway,{(Norway,11,2020,46,15184)},{(Norway,4,2020,176,3441)},7)
(Jersey,{(Jersey,11,2020,0,354)},{(Jersey,4,2020,19,223)},7)
(Mali,{(Mali,11,2020,13,1114)},{(Mali,5,2020,51,768)},6)
(Cayman_Islands,{(Cayman_Islands,5,2020,0,68)},{(Cayman_Islands,11,2020,1,34)},6)
(Andorra,{(Andorra,10,2020,22,2699)},{(Andorra,4,2020,34,373)},6)
(Belgium,{(Belgium,10,2020,1657,320023)},{(Belgium,4,2020,6609,34658)},6)
(Netherlands,{(Netherlands,10,2020,951,223554)},{(Netherlands,4,2020,3847,27052)},6)
(Japan,{(Japan,11,2020,364,46368)},{(Japan,5,2020,476,2763)},6)
(Northern_Mariana_Islands,{(Northern_Mariana_Islands,10,2020,0,22)},{(Northern_Mariana_Islands,4,2020,2,12)},6
```

Analysis 3

My third proposed analysis was to find the sequence in which the COVID-19 propagated around the world; for this, I wanted to list the dates of each country's first confirmed cases.

Analysis 3 results

We can see China is the first country where confirm case appeared this makes sense since it was in Wuhan, where the pandemic's epicentre took place. The virus then propagated mostly in China's surrounding areas (Thailand, Japan, South Korea, Taiwan, Vietnam, Singapore, Nepal etc...) It is interesting to find out that Mexico was the third country to have confirmed cases since Latinos seem to be hit by the pandemic much later. In contrast, the last lands to be hit by the virus were remote islands; this was also pointed out by analysis one. I found most amusing out of the analysis to see how Thailand is the second country in this list but show to be of the least affected; perhaps we can learn from them and their culture on containing a virus in a future.

```
(2019-12-31T00:00:00.000-05:00,27,China)
(2020-01-13T00:00:00.000-05:00,1,Thailand)
(2020-01-14T00:00:00.000-05:00,1,Mexico)
(2020-01-15T00:00:00.000-05:00,1,Japan)
(2020-01-20T00:00:00.000-05:00,1,South_Korea)
(2020-01-21T00:00:00.000-05:00,1,Taiwan)
(2020-01-21T00:00:00.000-05:00,1,United_States_of_America)
(2020-01-24T00:00:00.000-05:00,2,Vietnam)
(2020-01-24T00:00:00.000-05:00,3,Singapore)
(2020-01-25T00:00:00.000-05:00,1,Nepal)
(2020-01-25T00:00:00.000-05:00,3,France)
(2020-01-25T00:00:00.000-05:00,1,Australia)
(2020-01-25T00:00:00.000-05:00,3,Malaysia)
(2020-01-26T00:00:00.000-05:00,1,Canada)
(2020-01-27T00:00:00.000-05:00,1,United_Arab_Emirates)
(2020-01-28T00:00:00.000-05:00,1,Germany)
(2020-01-28T00:00:00.000-05:00,1,Cambodia)
(2020-01-28T00:00:00.000-05:00,1,Sri_Lanka)
(2020-01-30T00:00:00.000-05:00,1,Finland)
(2020-01-30T00:00:00.000-05:00,1,India)
(2020-01-30T00:00:00.000-05:00,1,Philippines)
(2020-01-31T00:00:00.000-05:00,3,Italy)
```

```
(2020-10-16T00:00:00.000-04:00,2,Solomon_Islands)
(2020-10-17T00:00:00.000-04:00,1,Wallis_and_Futuna)
(2020-10-29T00:00:00.000-04:00,1,Marshall_Islands)
(2020-11-11T00:00:00.000-05:00,1,Vanuatu)
```

Bonus

I continue my investigation on Thailand and found this interesting video I would like to share about how they are handling this pandemic. In the video, Elephants can be seen giving face masks for free to the population, which is amusing to watch. The footage also confirms my analysis in that Thailand was one fo the first countries to be hit but remained strong with their anti-outbreak measures.

https://www.youtube.com/watch?v=aYz2PE8tuOI&ab_channel=SouthChinaMorningPost