Resident in Europe who are part-time employed and their participation in tourism for personal purposes.

1st Richard Saldanha
Data Intensive Architectures (DIA)
National College of Ireland
Dublin, IRELAND
x18183034@student.ncirl.ie

***Abstract*—The aim of this project is to understand the behaviour of citizens of different countries in Europe who are part-time employed and their participation in tourism for personal purposes using MapReduce technique in Java programming language and by performing Pearson Coefficient of correlation in Python programming language.From the results achieved we found an inversely correlated correlation between the total population of the country who are part time employed and their participation in tourism for a particular country and year.It was quite surprising to observe that when the total percentage of part time employed residents in a particular country and year is low in those countries and year the total participating in tourism for personal purposes was high and a vice versa scenario was observed.From our analysis we also that there are many countries in a particular time phrase where there are no part time employed citizens and no participation in tourism for personal purposes like countries Belgium and Austria in the year 1996.There are valuable insights drawn during the analysis which can be found in the result section.**

***Index Terms*—Part Time Job, Participation in Tourism, Hadoop, MapReduce, ReduceSideJoin, Python, Java.**

I. Introduction

Do citizens in different countries in Europe who are part time employed like to travel within the country, to different countries or both for personal purposes with what they earn as it can have an impact on the overall growth and development of the country? In 2018, it was observed that roughly around 40 million of total of men and women population are part time employed, there could have been several reason like no proper education support, illness, family support at early age, could not find a full time job in the profession of choice and many more reasons could be there for their part time employment.The hypothesis which I am trying to achieve is that people who are doing a part time job because they could not find a full time job, spend their money wisely as they have limited source of income and would rather not invest their money on tourism for personal purposes as things are getting expensive day by day and spending money on leisure activities would be a bad idea as to sustain and pull out on expenses with a salary of a part time is difficult.

II. Data

1) The datasets for this project are taken from Eurostat-Your key to European statistics data and are licensed as open source.
2) The first dataset is about the main reason why residents in different countries in Europe do part-time job, there could be several reasons but, in this project, we have considered that they could not find a full-time job as the main reason[1].The geographical coverage are over a couple of countries in Europe. There are 24 age class of Females and Males under observation right from the age of 15 years to 75 years.The time period under observation is from 1994 to 2018.Thus, the csv file consists of 7 columns which are TIME, GEO, SEX, AGE, REASON, UNIT and Value with 22800 records.For more information about the dataset kindly refer to the meta data section which describes the codes used in the dataset.
3) The second dataset is about the number of trips taken by residents in different countries in Europe for personal purposes in a year[2].Depending upon personal choices and preferences the citizens may opt to travel within the country,may travel to other countries, or plan on something like travel to nearby places within a country and then may a trip to other country.Since many countries in Europe are famous holiday destinations there are residents who may travel to different countries of the world.The duration of travel could be for a single night or over,From 1 to 3 nights or even for 4 nights or over[2].The data is collected and observed from different geographical countries in Europe. Thus, the csv file consists of 6 columns which are TIME, GEO, UNIT, DURATION, PARTNER and Value consisting of 10800 number of records. For more information about the codes used in the dataset kindly refer to the meta data section.
4) The reason why I have chosen this topic for my project is to understand the behaviour of citizens in Europe having a part time job show interest towards going for a tour to different places either domestically or internationally in a year depending upon their preferences and for personal purposes which can tell us about their spending towards tourism which can have an impact on the GDP of the country as in most countries in Europe the main source of economy is tourism.

III. Methodology

1) The cleaning of the two datasets are done using python programming language.
2) We have made use of pandas package for data cleaning, manipulation and analysis of semi structured data like cvs and also we have made use of numpy package for handling and dealing with missing values in the csv files.
3) First we read the two csv files from our local system which are mainreasonforparttime.csv and participationintourismforpersonalpurposes.csv using read_csv file in pandas and store them as dataframes.
4) Once we have read the csv files then we check for any duplicate values in the file and we drop the duplicate values using the drop_duplicates command.
5) During the analysis we found that in the Value column there are missing values in both the csv files which is denoted by colon(:) so first we replace them with NaN values using the replace(':', np.nan, inplace = True)
6) Once, the values are converted to NaN then we fill the values with zero using fillna(0) command.
7) The datasets are now cleaned which means we have handled the missing values and there is no loss of data.
8) We then create a key which is made by contacting two columns TIME and GEO which is country_period which will act as a primary key in one table and foreign key in another table.
9) We then rename the column headers according to proper naming convention.
10) The ready dataframes are written back to csv file using the .to_csv command and the two new csv files are ready_mainreasonforparttime.csv and ready_participationintourismforpersonalpurpose.csv

IV. Implementation and Architecture

1) This project is implemented on Amazon Elastic Cloud Compute (EC2) service using Ubuntu Server 18.04 LTS (HVM), SSD Volume Type - ami-02df9ea15c1778c9c (64-bit x86) / ami-07a3c7461cc82f8ff (64-bit Arm) as the Amazon machine image(AMI).
2) The java version installed is 11.0.4 and the latest version of the Hadoop environment which is 3.1.3.
3) In this project I have implemented Hadoop MapReduce using Reduce Side Join.
4) In the reduce side join, "the reducer is responsible for performing the join operation"[5].
5) The reason why I have considered using Reduce Side join over the map side join is because it is simpler and easier to implement[5].
6) In Reduce Side Join "Mapper reads the input data which are to be combined based on common column or join key.The mapper processes the input and adds a tag to the input to distinguish the input belonging from different sources or data sets or databases.The mapper outputs the intermediate key-value pair where the key is nothing but the join key.After the sorting and shuffling phase, a key and the list of values is generated for the reducer. Now, the reducer joins the values present in the list with the key to give the final aggregated output[5]".
7) In the reduce side join sorting and shuffling phase sends the values having identical keys in our case the "country_period" to the same reducer and therefore, by default, the data is organized for us[5].
8) observe the **Fig. 1** which explains about the two mappers which are ParttimeMapper and TourismMapper.The key is country_period ("parts[7]") and we want "value in percent" which is "parts[6]". In tourismMapper the key is country_period which is at ("parts[5]") and we want the "value in percent" which is parts[4]".So the output in ParttimeMapper would be [BE2010,PARTIME 3.5], [AT2010, PARTTIME 4.5] etc.The output in TourismMapper would be [AT2009,TOURISM 7.8],[BE1995, TOURISM 3.4] etc.
9) In the next phase sorting and shuffling will take place "The sorting and shuffling phase will generate an array list of values corresponding to each key. In other words, it will put together all the values corresponding to each unique key in the intermediate key-value pair"[5].
10) In the reducer phase, perform the operation as it is highlighted to calculate the total of part time and total tourism and so the output would look like [AT1995, 19.5, 3.5],[BE2000, 4.5, 10.75].
11) Finally I will write the output to the outputfolder which is outDiaProjectMRJoin in the HDFS and a part-r-00000 would be created which will hold the output of the Hadoop MapReduce Reduce-Side Join.
12) **Fig. 3** Explains the complete process flow implementation of Hadoop MapReduce using Reduce-Side Join as the type of join operation.
13) The correlation.csv file is read using the pandas.read_csv file into a DataFrame and the quality of the data is handled by dropping the missing values and the labels i.e column headers with proper naming conventions is added so that there is a clear understanding of the dataset and the output generated would be accurate.
14) We perform Pearson Coefficient of correlation in Python to measure the strength of a linear association between the two variables total_parttime and total_tourism of the population of countries from 1994 to 2018. **Fig. 4** illustrates the Pearson Coefficient correlation matrix.

V. Results

1) To understand the nature of the datasets let us observe the output of the Hadoop MapReduce using reduce side join. From the **Fig. 6**.It is evident that there is an inverse correlation between the total population doing a part time job and the annual rate of total population participating in tourism for personal purposes for that particular country and year.
2) The results obtained is quite surprising, let us consider the country Austria(AT) from 1994 to 2018 we observe that when the rate of the population doing a part time job is low the rate of the

```java
public static class ParttimeMapper extends Mapper < Object, Text, Text, Text >
{
        public void map(Object key, Text value, Context context)
        throws IOException,InterruptedException
        {

                String record = value.toString(); //Read each record
                String[] parts = record.split(","); // Parse csv file
                //System.out.println("Parttime"+parts);
                context.write(new Text(parts[7]), new Text("PARTTIME@" + parts
        }
}


public static class TourismMapper extends Mapper < Object, Text, Text, Text >
{
        public void map(Object key, Text value, Context context)
        throws IOException,InterruptedException
        {
            String record = value.toString(); // Read each record
            String[] parts = record.split(","); //Parse csv file

            //System.out.println("Tourism"+parts[4]);
            context.write(new Text(parts[5]), new Text("TOURISM@" + parts[4]))
        }
}
```

Fig. 1. Mappers for Part time and Tourism

```java
public static class DiaProjectReduceJoinReducer extends Reducer <Text, Text, Text,
{
        public void reduce(Text key, Iterable<Text> values, Context context)
        throws IOException, InterruptedException
        {
        double totalparttime=0.0;
        double totaltourism=0.0;

        for (Text t : values)
        {
            String parts[] = t.toString().split("@");
            if (parts[0].equals("PARTTIME")& parts[1] != null)
             {

                totalparttime = Float.parseFloat(parts[1]);
             }
            else if(parts[0].equals("TOURISM")& parts[1] != null)
             {
                totaltourism = Float.parseFloat(parts[1]);
             }
        }
        String str = String.format("%f  %f",totalparttime,totaltourism);
        context.write(new Text(key), new Text(str));
        }
}
```

Fig. 2. Reducer

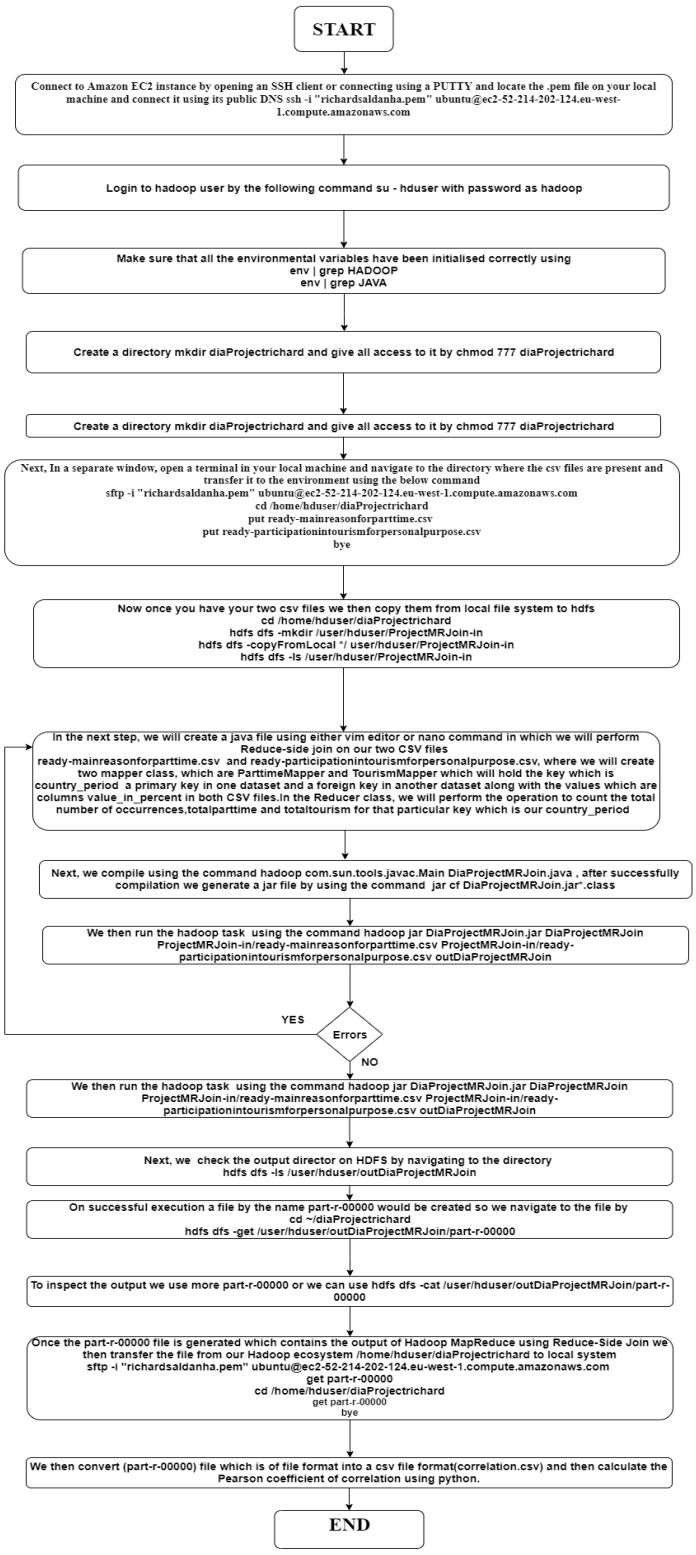

Fig. 3. Hadoop MapReduce using Reduce-Side Join

	total_parttime	total_tourism
total_parttime	1.000000	-0.204834
total_tourism	-0.204834	1.000000

Fig. 4. Output: Pearson Coefficient of Correlation Matrix

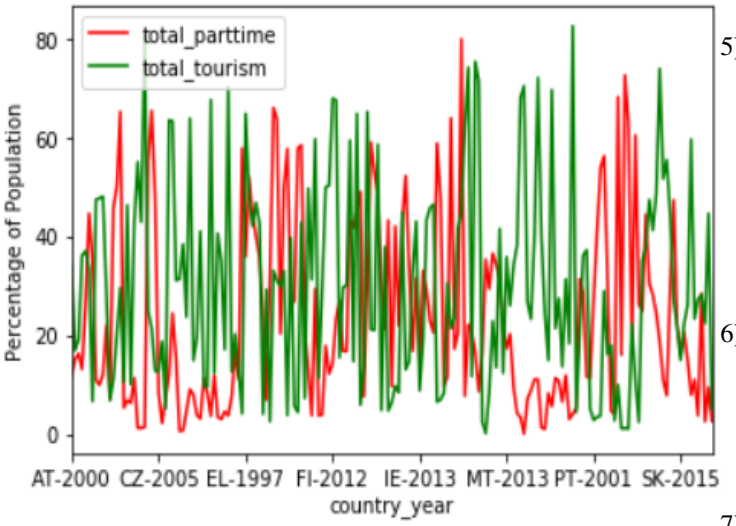

Fig. 5. Line Graph comparison between total part time and total tourism

population participating in tourism for personal purposes was high and vice versa.Considering the country Austria(AT) we observe that in the year 1994,1995,1996,2003,2005,2006,2007,2008,2010 the population of the country did not do a part time job and also there was no involvement in tourism for personal purposes.In 1997,1998, 2002,2004,2014,2018 the citizens of the country were doing a part time job but there was no involvement in tourism for personal purposes.

3) On comparing with the results of analysis obtained in Belgium(BE) we observe that in the year 1996 the residents were neither doing a part time job nor there was an interest in tourism for personal purposes.Like Austria(AT) it was also observed in Belgium(BE) that there is an inversely proportional relationship between the total part time employed residents and their participation in tourism for personal purposes.It was also observed that there are scenarios like in 1994, 1995, 1997, 2001, 2002, 2006 to 2007 the residents were doing a part time job but the percentage of the population participation in tourism for personal purposes like going on a vacation or a business trips was zero.
4) In the year 2009, residents who were part time employed in Austria and Belgium, the highest percentage observed was in Austria at a rate of 15.3 percent but it was surprising that citizens in Belgium participated more in tourism for their personal purposes compared to citizens of Austria, the observed rate was found to be 47.59 percent which is greater than the tourism percentage in Austria.In the year 2017, yet again the percentage of the population having a part time job was highest in Austria compared to Belgium, but it was astonishing to find out that the Austrian's had participated more in Tourism for their personal purposes despite doing part time jobs as compared to citizens of Belgium, the observed rate was found to be 36.29 percent.
5) In the similar manner, we observe an inverse correlation between total population having a part-time job and total participation in tourism for personal purposes measured in units of percentage for the remaining 37 countries in Europe from a period of 1994 to 2018 as can have a comparison between the countries like we did for Austria and Belgium to draw valuable insights from the output obtained through Hadoop MapReduce doing Reduce Side Join.
6) The expected results of the analysis was that people who are doing a part time job because they could not find a full time job, spend their money wisely as they have limited source of income and would rather not invest their money on tourism for personal purposes and we would observe a direct correlation between them.
7) From the Pearson Coefficient of Correlation Matrix in **Fig. 4** we can observe that the two variables total_parttime and total_tourism are inversely correlated with each other.A negative sign indicates there is a negative relationship between the two variables.For perfectly correlated variables the value is 1 which would be the diagonal elements.The coefficient of correlation using Pearson rank obtained is -0.204834which indicates higher the total part time population for a particular year and country lower would be the total participation in tourism for personal purposes in that year and country and vice versa.
8) We observe the figure **Fig. 5** which projects the comparison between total part time employed citizens and the total participation in tourism for personal purposes for a particular country and year.The countries and year under observations are AT(Austria)-2000, ,CZ(Czechia)-2005, EL(Greece)-1997, FI(Finland)-2012, IE(Ireland)-2013, MT(Malta)-2013,(PT)Portugal-2001,Slovakia(sk)-2015.From the graph we can observe that there is an upward downward trend in the percentage of total part time employed residents and the total participation in tourism for personal purposes which means that with an in increase in the percentage of total part time employed rate there would be a decrease in the total population

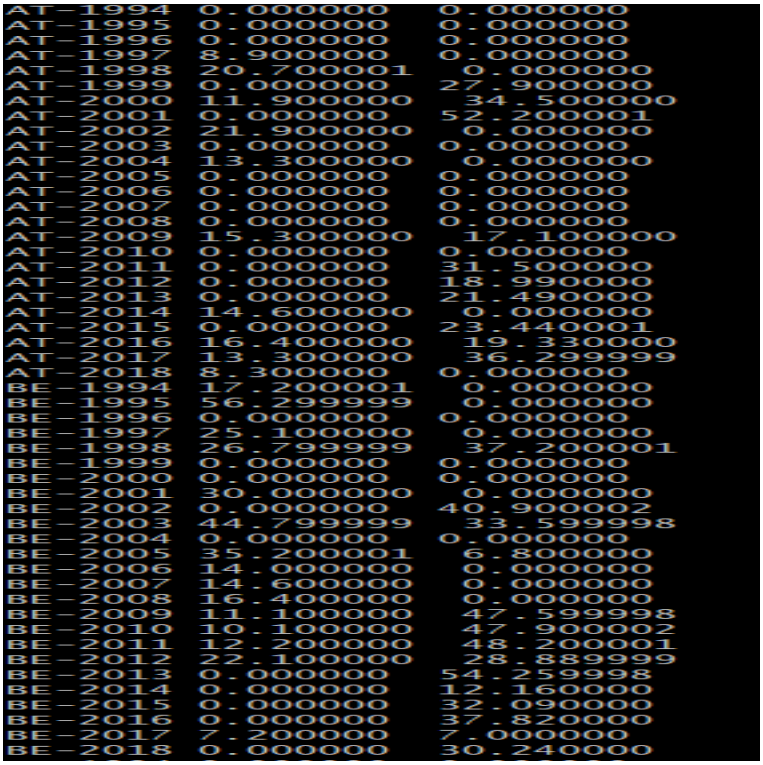

Fig. 6. Output: Reduce-Side Join

in participation in tourism for personal purpose for that particular country and year.From the graph it is evident that between CZ-2005 and EL-1997,MT-2013 and PT-2001 the total part time employed was roughly below 25 % while the total participation in tourism for personal purpose rate was high above 40 % , the highest rate of total tourism participation for personal purpose was highest around CZ-2005 and PT-2001 and the highest part time employed residents were in IE-2013 with moderately low participation rate in participation of tourism for personal purposes.

VI. Conclusions and Future Work

1) I learnt from my project that there is an inverse relation between the people doing a part time job and their participation in tourism for personal purposes
2) An interesting aspect of the analysis observed via the results of Hadoop MapReduce performing Reduce Side Join, the Pearson correlation coefficient matrix and the line graph showed us that there is an inversely proportional relationship between the total population who are part time employed and participation in tourism for personal purposes means with an increase in total Part time rate the total tourism rate would decrease and vice versa scenario would be observed in a particular country and year which is quite surprising because residents who are part time employed would generally not splurge their earnings on tour for their personal purposes but rather invest in other activities to manage their livelihood due to less income compared to full time employed residents so a direct relationship was expected between the total part time employed and total participation in tourism for personal purpose for a particular country and year.
In this project I had done a comparison between the resident who are part time employed and their participation in tourism but in the next project I would try find out a correlation between Part time employed residents in Europe and the crime rates in that particular area, with an hypothesis that does people who are doing part time job have a criminal mindset and commit crimes.
If there were more time permitted I would be interested to find out which age groups who are part time employed participate more in tourism for their personal purposes for each year and countries under observation.

VII. Metadata

In The first dataset the indicators are could not find a full-time job (INVPT) The geographical coverage are (EU28)-European Union-28 countries, (EU15)-European Union-15 countries from 1995 to 2004,EA19-Euro area consisting of 19 countries, (BE)-Belgium, (BG)-Bulgaria, (CZ)-Czechia, (DK)-Denmark, (DE)-Germany(until 1990 former territory of the FRG), (EE)-Estonia, (IE)-Ireland, (EL)-Greece, (ES)-Spain, (FR)-France, (HR)-Croatia, (IT)-Italy, (CY)-Cyprus,(LV)-Latvia, (LT)-Lithuania, (LU)-Luxembourg, (HU)-Hungary,(MT)-Malta,(NL)-Netherlands,(AT)-Austria,(PL)-Poland,(RO)-Romania,(SI)-Slovenia,(FI)-Finland,(SE)-Sweden,(UK)-United Kingdom,(SK)-Slovakia (IS)-Iceland,(NO)-Norway,(CH)-Switzerland,(ME)-Montenegro,(MK)-North Macedonia, (RS)-Serbia and (TR)-Turkey.A Total of Males and Females population is taken under observation which is coded as (T) and the unit of measurement is Percentage coded as (PC).The age split categories are (Y15-19)-15 to 19 years, (Y15-24)-15 to 24 years, (Y15-29)-15 to 29 years, (Y15-39)-15 to 39 years, (Y15-59)-15 to 59 years, (Y15-64)-15 to 64 years, (Y15-74)-15 to 74 years,(Y_GE15) 15 years or over, (Y20-64)- 20 to 64 years,(Y25-49)-25 to 49 years, (Y25-59)-25 to 59 years, (Y25-64)- 25 to 64 years, (Y25-74)-25 to 74 years, (Y_GE25)-25 year or over, (Y40-59)-40 to59 years,(Y40-64)-40 to 64 years,(Y50-59)-50 to 59 years,(Y50-64)-50 to 64 years, (Y50-74)-50 to 74 years, (Y_GE50) -50 years or over, (Y55-64)-55 to 64 years, (Y55-74)-55 to 74 years, (Y_GE65)-65 years or over, (Y_GE75)-75 years or over.For more information on the dataset kindly refer to the metadata “LFS series-detailed annual survey results (Ifsa)[3].”

2) In the second dataset is the time period for this dataset under observation is from 1994 to 2018 which is measured in Percentage of total population (PC_POP). Depending upon the interest the citizens may opt to travel within the country (DOM)Domestically,(OUT)

Outbound-Traveling to different countries or a combination of both Domestic and outbound (DOT),Travel to different countries of the world coded as (WORLD).The duration of travel (N_GE1)-1 night or over, (N1-3)-From 1 to 3 nights or even for 4 nights or over (N_GE4).The data collected is from (EU28)-European Union-28 countries, (EU27)-European Union- 27 countries from 2007 to 2013, (EU25)- European Union – 25 countries from 2004 to 2006, (EA)-Euro area (EA11-2000,EA12-2006,EA13-2007,EA15-2008,EA16-2010,EA17-2013,EA18-2014,EA19), (BE)-Belgium, (BG)-Bulgaria, (CZ)-Czechia, (DK)-Denmark, (DE)-Germany(until 1990 former territory of the FRG), (EE)-Estonia, (IE)-Ireland, (EL)-Greece, (ES)-Spain, (FR)-France, (HR)-Croatia, (IT)-Italy, (CY)-Cyprus,(LV)-Latvia, (LT)-Lithuania, (LU)-Luxembourg, (HU)-Hungary, (MT)-Malta,(NL)-Netherlands,(AT)-Austria,(PL)-Poland,(PT)-Portugal, (RO)-Romania,(SI)-Slovenia,(FI)-Finland,(SE)-Sweden,(UK)-United Kingdom, (IS)-Iceland,(NO)-Norway,(CH)-Switzerland,(ME)-Montenegro and (MK)-North Macedonia. For more information on the dataset kindly refer to the metadata on “Annual data on trips of EU residents(tour_dem)[4]”.

REFERENCES

[1] Ec.europa.eu. (2019). Main reason for part-time employment - Distributions by sex and age(%) - Eurostat. [online] Available at: https://ec.europa.eu/eurostat/web/products-datasets/product?code=LFSA_EPGAR [Accessed 2 Dec. 2019].

[2] ”Ec.europa.eu. (2019). Participation in tourism for personal purposes - Eurostat. [online] Available at: https://ec.europa.eu/eurostat/web/products-datasets/-/tour_dem/_totot [Accessed 2 Dec. 2019].

[3] ”LFS series - detailed annual survey results (lfsa)”, Ec.europa.eu, 2019. [Online]. Available: https://ec.europa.eu/eurostat/cache/metadata/en/lfsa_esms.htm. [Accessed: 05- Dec- 2019].

[4] ”Annual data on trips of EU residents (tour_dem)”, Ec.europa.eu, 2019. [Online]. Available: https://ec.europa.eu/eurostat/cache/metadata/en/tour_dem_esms.htm. [Accessed: 05- Dec- 2019].

[5] A. Bakshi, ”MapReduce Example — Reduce Side Join MapReduce Example — Edureka”, Edureka, 2019. [Online]. Available: https://www.edureka.co/blog/mapreduce-example-reduce-side-join/. [Accessed: 06- Dec- 2019].