Jurnal KomtekInfo , Vol.7 No.1 Tahun 2020
e-ISSN : 2502-8758 || p-ISSN : 2356-0010
DOI : 10.35134/komtekinfo.v7i1
Page : 01-14

Data Preprocessing: Case Study on monthly number of visitors to Taiwan by their residence and purpose

Muhammad Reza Putra [1] , Azuraliza Abu Bakar [2]
Universitas Putra Indonesia YPTK Padang[1], Universiti Kebangsaan Malaysia[2]
E-mail: Mhd.rezaputra@gmail.com

Abstract

This paper will explain in details on data reports preliminary on dataset, how the pre-processing data mainly for data cleaning and reduction process applied to a dataset. The dataset that will be used is number of visitors to Taiwan by their residence and purpose.Dataset which is obtained based on kaggle, findings from Scraped from Taiwan Tourism Bureau. The surveys have been carried out using Foreign visitor data covers all foreign visitors directly arrived in Taiwan through the airports, ports and land.

Keyword: Data preliminary reports, data pre-processing, data mining, data reduction, statistical method, number of visitor to Taiwan by residence and purpose

1. Introduction

The tourism industry is an increasingly important national industry for Taiwan. Government policymakers and business managers pay close attention to the development of the tourism industry. Accurate forecasts of demand for international tourism are important to effectively promote tourism and to allocate sufficient resources for operations, marketing, investment, and financial planning for the Taiwanese tourism industry. Although forecasting demand is vital to all industrial planning, forecasting is particularly crucial in the tourism industry because tourism products and services are inherently perishable.[1]

Tourist is is any visitor according to the definition above, staying at least 24 hours, but not more than 12 (twelve) months, in the place visited, with the intention of visiting, among others for the purpose of :[2]
A. Personal: pleasure, recreation, visiting friends and relatives, study and training, helath and medical care, spors, religion/pilgrimages, shopping, transit, etc.
B. Business and professional: attending meetings, conferences or congresses, trade fairs and exhibitions, concerts, shows, etc.

2. Literatur Review

The dataset that will be used is number of visitors to Taiwan by their residence and purpose, 2011-2018 dataset which is obtained based on the findings from kaggle where orginal data from Tourism Bureau[3]. MOTC Republic of China. The surveys have been carried out using Foreign visitor data covers all foreign visitors directly arrived in Taiwan through the airports, ports and land. Based on purpose of visitor.[4]

Submitted : 18 Desember, 2019 Accepted : 7 Januari 2020 Published : 1 Februari 2020
DOI : 10.35134/komtekinfo.v7i1

2.1. List of tools

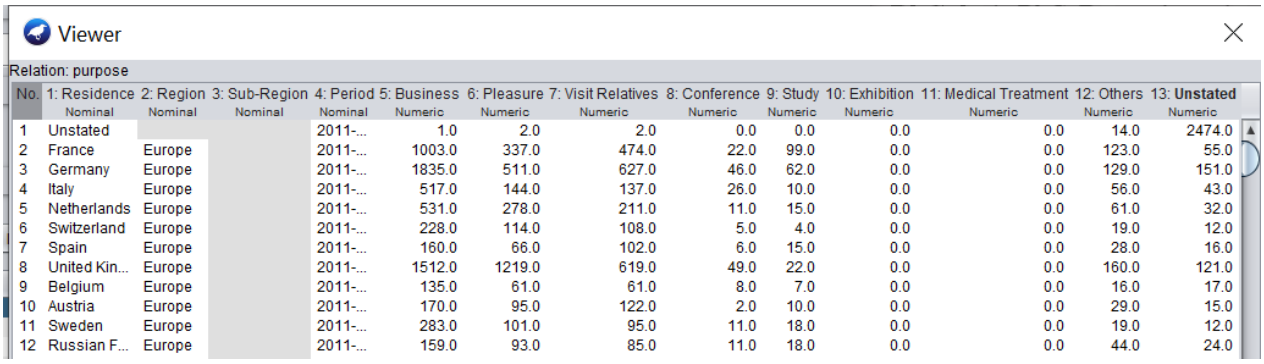
Table 1. List of tools

Tools	Function
Microsoft Excel 365	- Examine the instances horizontally and vertically to find empty column and instances - Data Reduction - Data Transformation - Data examination to find incomplete, noisy data - Handling outliers - To convert file type excel (.xlsx) to a comma-separated values (CSV) file. CSV is a delimited text file that uses a comma to separate values. A CSV file stores tabular data (numbers and text) in plain text
Weka 3.8	- To convert file type csv to ARFF format. This is because Weka prefers to load data in the ARFF format. ARFF is an acronym that stands for Attribute-Relation File Format. It is an extension of the CSV file format where a header is used that provides metadata about the data types in the columns - Data monitoring, visualization - Statistic of data - missing value data - Data discretization - Handling outliers - Feature selection task - Analysis and classification task

The main objective of this research is to extract the survey data carried out using Foreign visitor data covers all foreign visitors directly arrived in Taiwan through the airports, ports and land. The purpose of this paper is to forecast the number of tourists in advance for the government of Taiwan so that the tourism department shall be prepared in advance to provide essential services to the forthcoming tourists.

2.2. Attributes and Data Quality Report

Using Weka 3 (Data Mining Software in Java), from the raw tourism arrival dataset, there are 13 attributes that consist same types of variables. Because of the data is time series data so the data only defide into one section.



Viewer

Relation: purpose

No.	1: Residence Nominal	2: Region Nominal	3: Sub-Region Nominal	4: Period Nominal	5: Business Numeric	6: Pleasure Numeric	7: Visit Relatives Numeric	8: Conference Numeric	9: Study Numeric	10: Exhibition Numeric	11: Medical Treatment Numeric	12: Others Numeric	13: **Unstated** Numeric
1	Unstated			2011-...	1.0	2.0	2.0	0.0	0.0	0.0	0.0	14.0	2474.0
2	France	Europe		2011-...	1003.0	337.0	474.0	22.0	99.0	0.0	0.0	123.0	55.0
3	Germany	Europe		2011-...	1835.0	511.0	627.0	46.0	62.0	0.0	0.0	129.0	151.0
4	Italy	Europe		2011-...	517.0	144.0	137.0	26.0	10.0	0.0	0.0	56.0	43.0
5	Netherlands	Europe		2011-...	531.0	278.0	211.0	11.0	15.0	0.0	0.0	61.0	32.0
6	Switzerland	Europe		2011-...	228.0	114.0	108.0	5.0	4.0	0.0	0.0	19.0	12.0
7	Spain	Europe		2011-...	160.0	66.0	102.0	6.0	15.0	0.0	0.0	28.0	16.0
8	United Kin...	Europe		2011-...	1512.0	1219.0	619.0	49.0	22.0	0.0	0.0	160.0	121.0
9	Belgium	Europe		2011-...	135.0	61.0	61.0	8.0	7.0	0.0	0.0	16.0	17.0
10	Austria	Europe		2011-...	170.0	95.0	122.0	2.0	10.0	0.0	0.0	29.0	15.0
11	Sweden	Europe		2011-...	283.0	101.0	95.0	11.0	18.0	0.0	0.0	19.0	12.0
12	Russian F...	Europe		2011-...	159.0	93.0	85.0	11.0	18.0	0.0	0.0	44.0	24.0

Picture 1 . Attributes and Data Quality

Table 2. List of attributes from raw dataset

Number	Atribute	Type	Description
1	Residence	Nominal	Country origin
2	Region	Nominal	an area, especially part of a country or the world having definable characteristics but not always fixed boundaries
3	Sub Region	Nominal	
4	Period	Nominal	Year and month of tourist arrival
5	Bussines	Numeric	Purpose of visit
6	Pleasure	Numeric	Purpose of Visit
7	Visit Relatives	Numeric	Purpose of Visit
8	Conference	Numeric	Purpose of Visit
9	Study	Numeric	Purpose of Visit
10	Exhibition	Numeric	Pupose of Visit
11	Medical Treatment	Numeric	Purpose of Visit
12	Others	Numeric	Purpose of Visit
13	Unstated	Numeric	Purpose of Visit

Viewer

Relation: purpose

No.	1: Residence	2: Region	3: Period	4: Business	5: Pleasure	6: Visit Relatives	7: Conference	8: Study	9: Exhibition	10: Medical Treatment	11: Others	12: **Unstated**
	Nominal	Nominal	Nominal	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric
1	Unstated		2011-...	1.0	2.0	2.0	0.0	0.0	0.0	0.0	14.0	2474.0
2	France	Europe	2011-...	1003.0	337.0	474.0	22.0	99.0	0.0	0.0	123.0	55.0
3	Germany	Europe	2011-...	1835.0	511.0	627.0	46.0	62.0	0.0	0.0	129.0	151.0
4	Italy	Europe	2011-...	517.0	144.0	137.0	26.0	10.0	0.0	0.0	56.0	43.0
5	Netherlands	Europe	2011-...	531.0	278.0	211.0	11.0	15.0	0.0	0.0	61.0	32.0
6	Switzerland	Europe	2011-...	228.0	114.0	108.0	5.0	4.0	0.0	0.0	19.0	12.0
7	Spain	Europe	2011-...	160.0	66.0	102.0	6.0	15.0	0.0	0.0	28.0	16.0

Picture 2 . List of attributes from new pre-process dataset

Table 2 . Attributes from new pre-process dataset

Number	Attribute	Type	Description
1	Residence	Nominal	Country origin
2	Region	Nominal	especially part of a country or the world having definable characteristics but not always fixed boundaries.
3	Period	Nominal	Year and month of tourist arrival
4	Bussines	Numeric	Purpose of visit
5	Pleasure	Numeric	Purpose of Visit
6	Visit Relatives	Numeric	Purpose of Visit
7	Conference	Numeric	Purpose of Visit
8	Study	Numeric	Purpose of Visit
9	Exhibition	Numeric	Pupose of Visit
10	Medical Treatment	Numeric	Purpose of Visit
11	Others	Numeric	Purpose of Visit
12	Unstated	Numeric	Purpose of Visit

We delete sub region so we can concentrate with usefull atribut. So now we only have 11 atribute. The data quality report for the new pre-process dataset are described in Table 3.

Feature	Count	% Miss	distinct	Min	Max	Mean	stddev
Bussiness	3,616	0	1785	0	28,448	1,757.584	4,033.97
Pleasure	3,616	0	1958	0	346,188	13,567.35	40,213.616
Visit Relatives	3,616	0	1417	0	20,029	955.84	2,218.159
Conference	3,616	0	651	0	3406	141.011	235.195
Study	3,616	0	581	0	8,600	144.054	472.228
Exhibition	3,616	0	251	0	1076	28	62.538
Medical Treatment	3,616	0	235	0	12,669	105.179	742.44
Others	3,616	0	1,859	3	64,051	2,915.196	6,966.295
Unstated	3,616	0	179	0	15,091	69.361	645.763

Table 3. Data quality report

2.3 Related Work

From the owner of the dataset that is Tourism Boureu MOTC A total of 836,594 visitors arrived in the Republic of China in September, 2018, increasing 16,491 or 2.01% from the 820,103 in September of last year. The arrivals included 507,761 foreign visitors and 328,833 Overseas Chinese. Compared with September of last year, the number of foreign visitors increased by 23,558 or 4.87%, and the number of Overseas Chinese visitors decreased by 7,067 or 2.10%. Daily arrivals in September averaged 27,886

I. **Main markets by residence**

1. Mainland China accounted for 218,125 or 26.07% of the total, down 2.92%, consisting of 2,982 foreign visitors, up 13.13%, and 215,143 Overseas Chinese, down 3.11%.

2. Hong Kong and Macao, 120,872 or 14.45%, down 0.52%, consisting of 8,944 foreign visitors, down 3.14%, and 111,928 Overseas Chinese, down 0.31%.

3. Japan accounted for 163,103 or 19.50% of the total, down 0.78%, consisting of 162,962 foreign visitors, down 0.81%, and 141 Overseas Chinese, up 39.60%.

4. Korea, 77,457 or 9.26%, up 7.56%, consisting of 77,053 foreign visitors, up 7.45%, and 404 Overseas Chinese, up 32.89%.

5. Southeast Asia, 166,508 or 19.90%, up 9.49%, consisting of 165,847 foreign visitors, up 9.52%, and 661 Overseas Chinese, up 2.64%.

6. U.S.A., 36,503 or 4.36%, up 1.02%, consisting of 36,193 foreign visitors, up 0.97%, and 310 Overseas Chinese, up 7.27%.

7. Australia and New Zealand, 8,525 or 1.02%, up 16.94%, consisting of 8,492 foreign visitors, up 16.99%, and 33 Overseas Chinese, up 6.45%.

8. Europe, 27,233 or 3.26%, up 12.91%, consisting of 27,197 foreign visitors, up 13.00%, and 36 Overseas Chinese, down 29.41%.

9. Other countries or regions accounted for 18,268 or 2.18%, up 2.20%, consisting of 18,091 foreign visitors, up 2.07%, and 177 Overseas Chinese, up 18.00%.

II. **Main markets by country of nationality**

1. There were 162,869 visitors from Japan accounting for 19.47% of the total. This was down 1,083, or 0.66%, from the number in the same month last year.

2. There were 77,637 visitors from Korea accounting for 9.28% of the total. This was up 5,645, or 7.84%, from the number in the same month last year.

3. There were 165,506 visitors from Southeast Asia accounting for 19.78% of the total. This was up 14,539, or 9.63%, from the number in the same month last year.

4. There were 37,592 visitors from U.S.A accounting for 4.49%% of the total. This was up 226, or 0.60%, from the number in the same month last year.

5. There were 9,815 visitors from Australia and New Zealand accounting for 1.17% of the total. This was up 981, or 11.10%, from the number in the same month last year.

6. There were 33,437 visitors from Europe accounting for 4.00% of the total. This was up 3,280, or 10.88%, from the number in the same month last year. Visitors from U.K., Germany, and France accounted for 0.96% (8,052 people), 0.72% (5,982 people), and 0.56% (4,673 people) of the total which were up 7.43%, 15.50%, and 15.61% from the number in the same month last year.

7. Other nationals accounted for 20,905 or 2.50%, down 30 or 0.14%.

8. Overseas Chinese visitors accounted for 328,833 or 39.31%, down 7,067 or 2.10%. Overseas Chinese visitors from Hong Kong and Macao accounted for 111,928, or 34.04%, of the total Overseas Chinese visitors.

III. **Main markets by type of visitor**

1. Mode of transportation and port of entry: There were 795,090 visitors, 95.04%, by air of whom 626,267, or 74.86%, through the Taiwan Taoyuan International Airport, 72,087, or 8.62%, through Kaohsiung Airport. Meanwhile, there were 41,504 visitors, 4.96%, by sea, 1,311 people, or 0.16%, through Kaohsiung port, 3,312 people, or 0.40%, through Keelung, 2,490 people, or 0.30%, through Taichung, 34,391 people, or 4.11%, through other ports.

2. Gender: Male visitors accounted for 47.51% (397,450 people) while female were 52.49% (439,144 people).

3. Age: There were 55,215 (6.60%) visitors under 19-year-old, 426,397 (50.97%) between 20 to 39-year-old, 258,058 (30.85%) between 40 to 59-year-old, and 96,924 (11.59%) over 59-year-old.

4. Purpose of visit: The main purposes are leisure, business, and relatives visiting with 556,312 (66.50%), 62,373 (7.46%), and 31,997 (3.82%) visitors.

Length of stay: Most of visitors, 168,577 people or 22.38%, stayed for 3 nights. Visitors staying 5 to 7 nights, 156,405 people or 20.76%, ranked second. Visitors staying 2 nights, 121,335 people or 16.11%, ranked third. The average length of stay of visitors leaving in September was 6.13 nights.

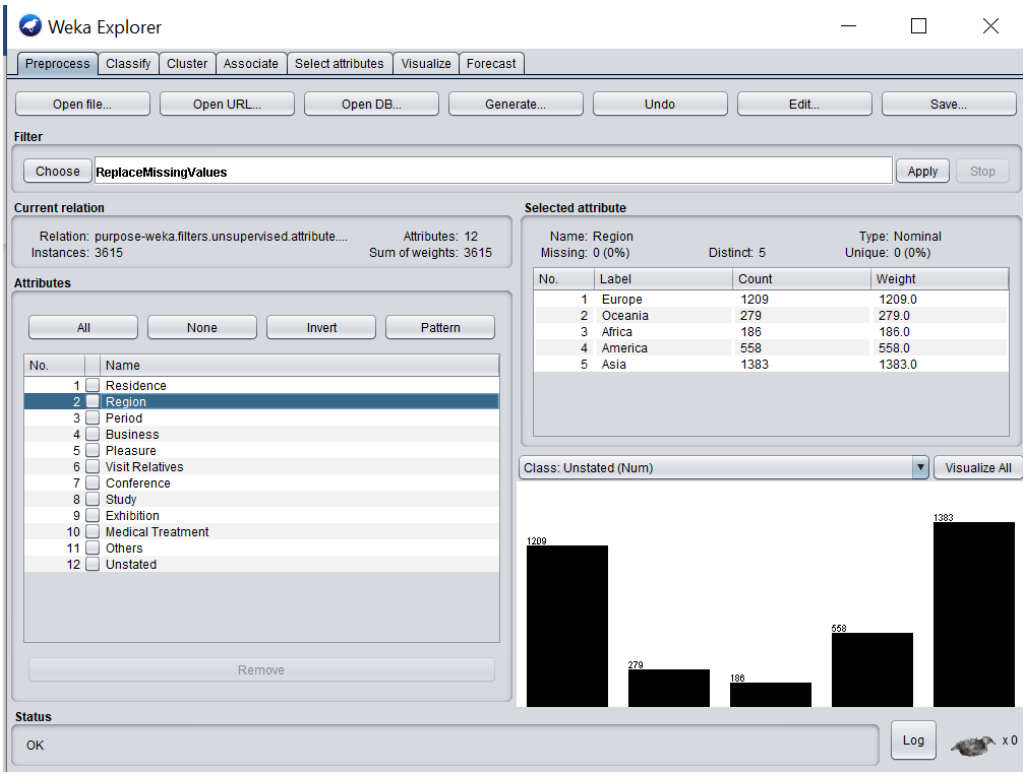
3. Methodology

3.1 Data Cleaning - Preliminary examination of data

Data cleaning, or data cleansing, is an important part of the process involved in preparing data for analysis. Data cleaning is a subset of data preparation, which also includes scoring tests, matching data files, selecting cases, and other tasks that are required to prepare data for analysis. Missing and erroneous data can pose a significant problem to the reliability and validity of study outcomes. Many problems can be avoided through careful survey and study design. During the study, watchful monitoring and data cleaning can catch problems while they can still be fixed. At the end of the study, multiple imputation procedures may be used for data that are truly irretrievable.

The problem of data may be caused by user entry errors, by corruption in transmission or storage in different location or same, or by different data dictionary definitions of similar entities during data integration. The raw data set need to be examine before well-planned pre-processing data task been deploy. Using Weka, the basic pattern can be produced and had shown significant pattern.

For the new dataset with 12 attributes, there are just only one attributes with missing value found. The attribute is region with 1 missing data. The missing value is replace using ReplaceMissingValue filter in Weka. The process is showed in figure 1 below



Picture 3. Data Cleaning

3.2 Methods-Prepocessing Technique used to each attribute in dataset

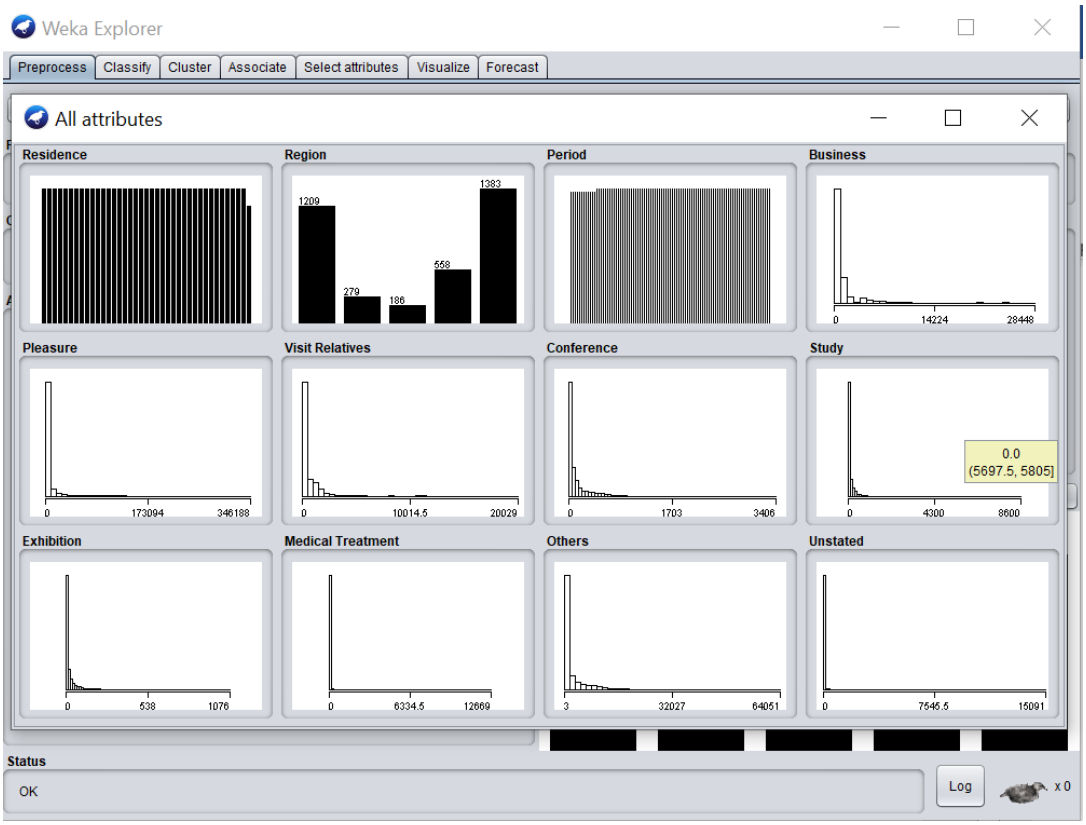
Before getting to the data cleaning process, it is important to identify the basic pattern of features in a dataset. From the features pattern in visualization before applying pre-processing data for attributes 1 to 16. We can see a few attributes that need to examine to ensure it's significant to ensure higher accuracy achievement during classification model task. In this project, using statistical method to ensure the attributes

and instances are sufficient and useful for modelling process while the other will be removed or retain. The description on each attributes on data preprocessing techniques is showed in table 4.

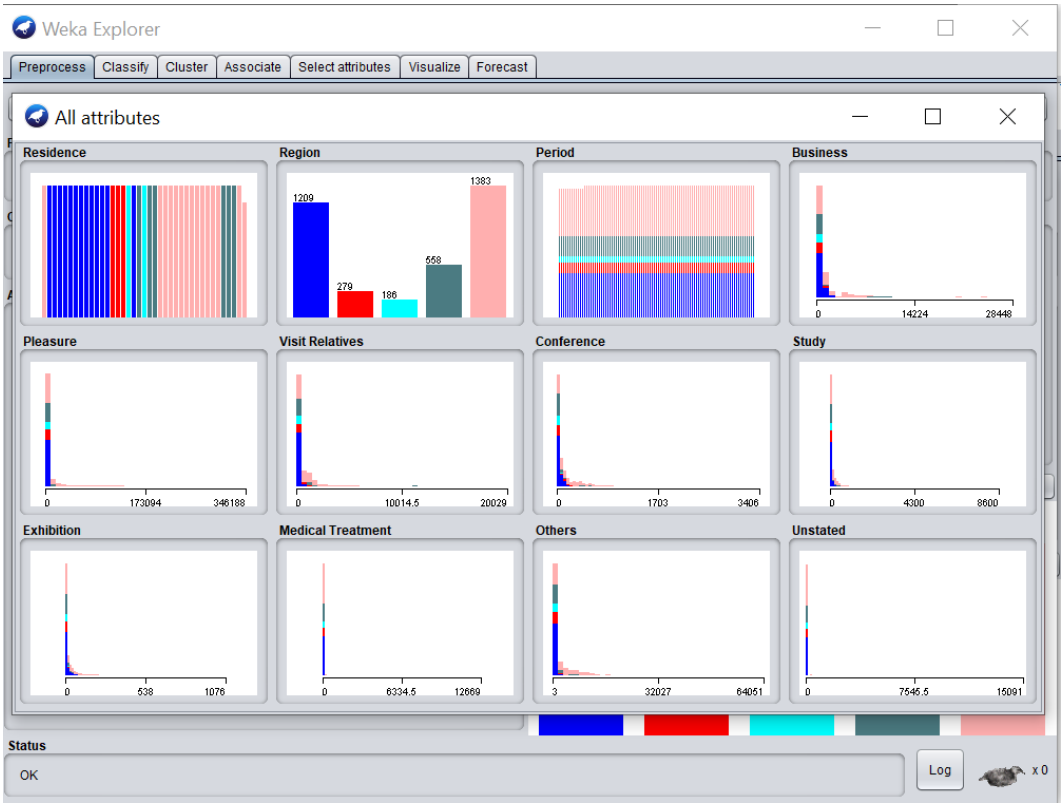
Table 4. description on each attributes on data preprocessing technique

Number	Atribute	Type	Description
1	Residence	Nominal	Country origin
2	Region	Nominal	an area, especially part of a country or the world having definable characteristics but not always fixed boundaries
3	Period	Nominal	Year and month of tourist arrival
4	Bussines	Numeric	Purpose of visit
5	Pleasure	Numeric	Purpose of Visit
6	Visit Relatives	Numeric	Purpose of Visit
7	Conference	Numeric	Purpose of Visit
8	Study	Numeric	Purpose of Visit
9	Exhibition	Numeric	Pupose of Visit
10	Medical Treatment	Numeric	Purpose of Visit
11	Others	Numeric	Purpose of Visit
12	Unstated	Numeric	Purpose of Visit

Attributes remove sub regionis been removed. Figure 2 before region become class. Figure 3 below shows all the features pattern in visualization after **region** is changed to nominal and become a class.

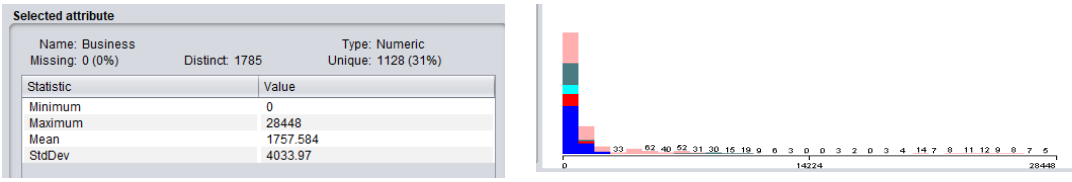


Picture 4. All Visualization for each attribute before region become class

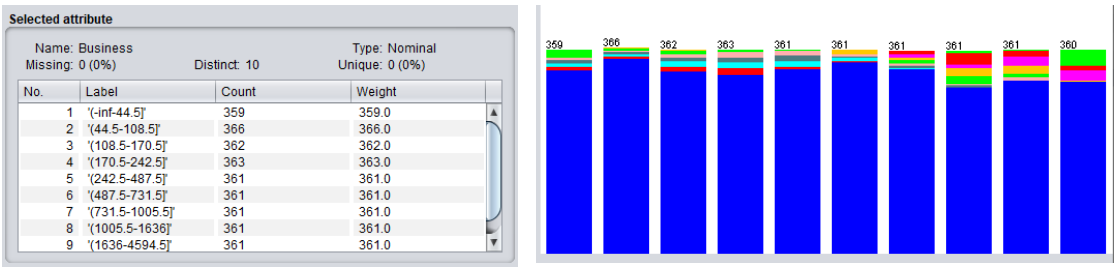


Picture 5. All Visualization for each attribute after region become class

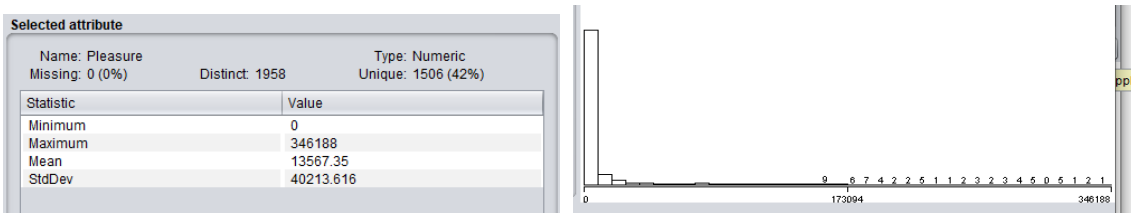
The graph for bussines, pleasure, visit relatives, conference, study, exhibition, medical treatment, others, unstated attribute shows it is not balance as shown in Figure 3 Attribute Umur(before). Using discretize technique, we use 10 bin to segregate the bussines, pleasure, visit relatives, conference, study, exhibition, medical treatment, others, unstated items. The result shows age attribute more equal and well-balanced.



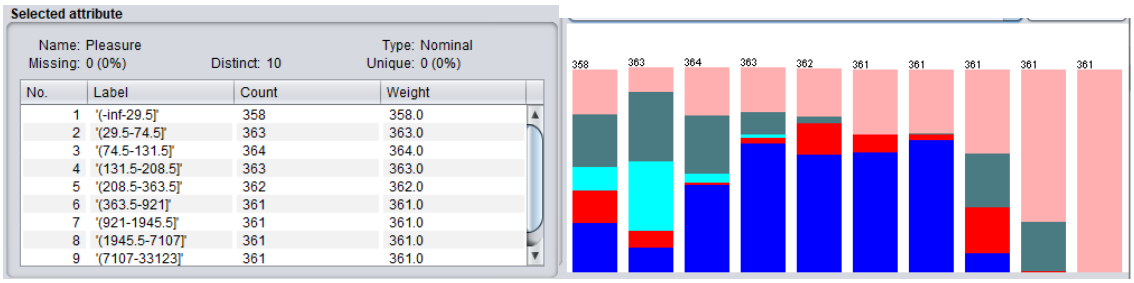
Picture 6. Attribute Bussiness Before



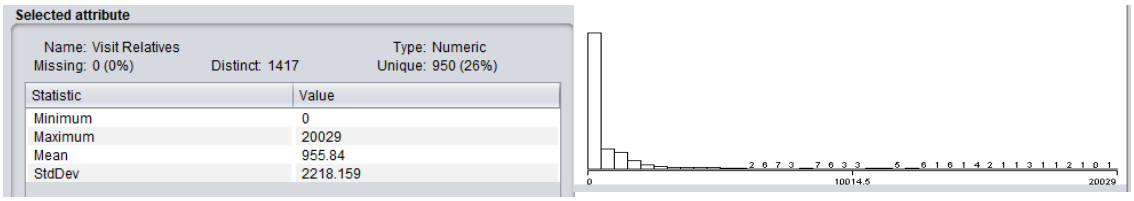
Picture 7. Attribute Bussiness After



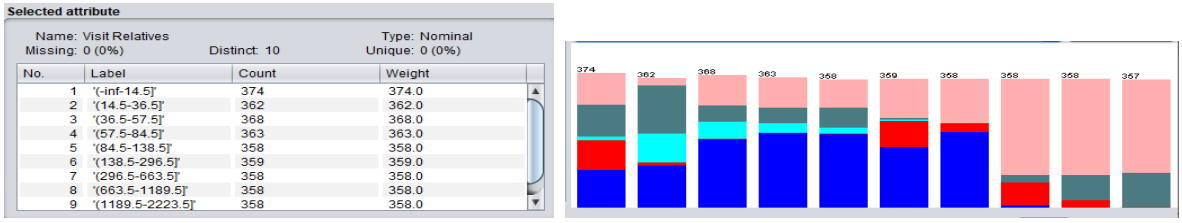
Picture 8. Attribute Pleasure Before



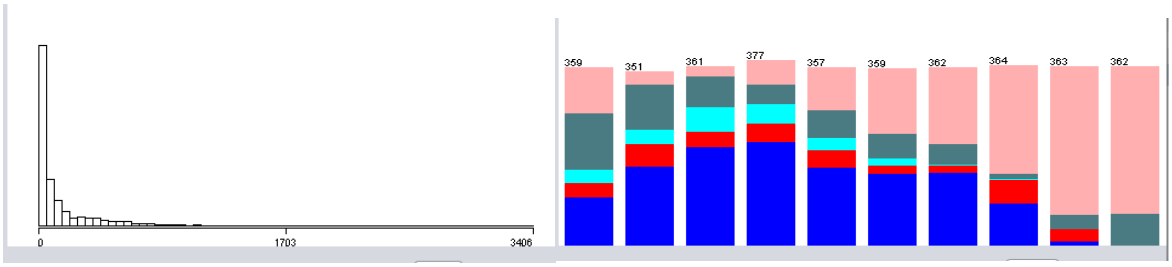
Picture 9. Attribute Pleasure After



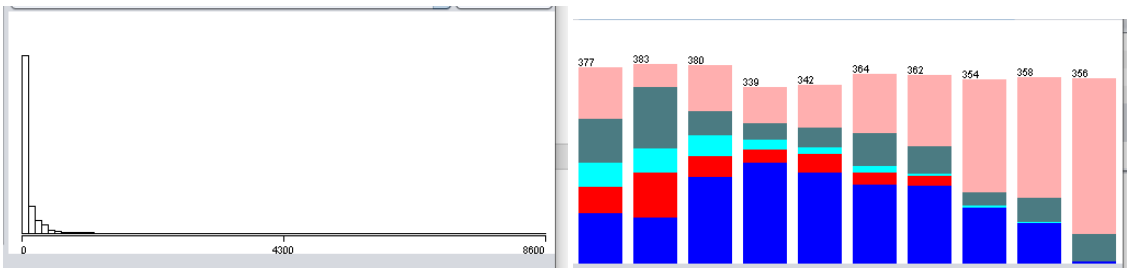
Picture 10. Attribute Visit Relatives Before



Picture 11. Attribute Visit Relatives After



Picture 12. Atributes Conference Before And After



Picture 13. Atributes Study Before And After

3.3 Feature Selection

We all may have faced this problem of identifying the related features from a set of data and removing the irrelevant or less important features with do not contribute much to our target variable in order to achieve better accuracy for our model. Feature Selection is one of the core concepts in machine learning which hugely impacts the performance of your model. The data features that you use to train your machine learning models have a huge influence on the performance you can achieve. Irrelevant or partially relevant features can negatively impact model performance. Feature selection and Data cleaning should be the first and most important step of your model designing.

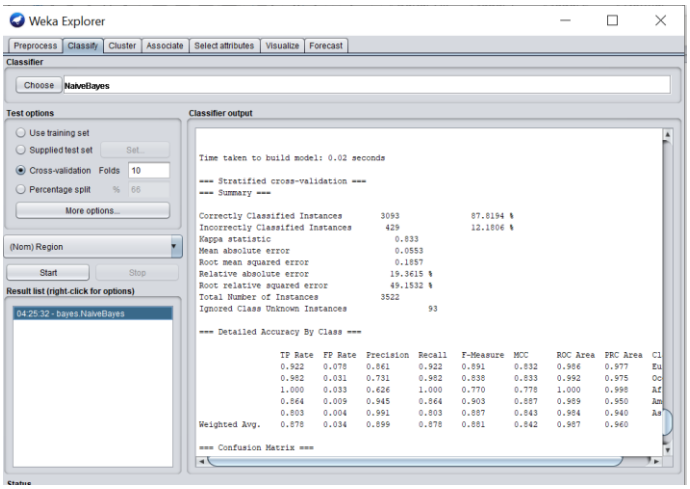
Feature selection method also helps modelling process become faster since size of the dataset has been reduced. In this dataset, as showed in figure 22 some attributes we can see that it doesn't have any correlation with expenditure exceed income prediction task.

```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (numeric): 12 Unstated):
        Correlation Ranking Filter
Ranked attributes:
 0.0841   2 Region
 0.0709   9 Exhibition
 0.0684  10 Medical Treatment
 0.0467   4 Business
 0.0442   5 Pleasure
 0.0432   7 Conference
 0.0427   6 Visit Relatives
 0.0401  11 Others
 0.0373   8 Study
 0.026    1 Residence
 0.0195   3 Period

Selected attributes: 2,9,10,4,5,7,6,11,8,1,3 : 11
```

Picture 14. Feature Selection using filter method

3.4 Evaluation

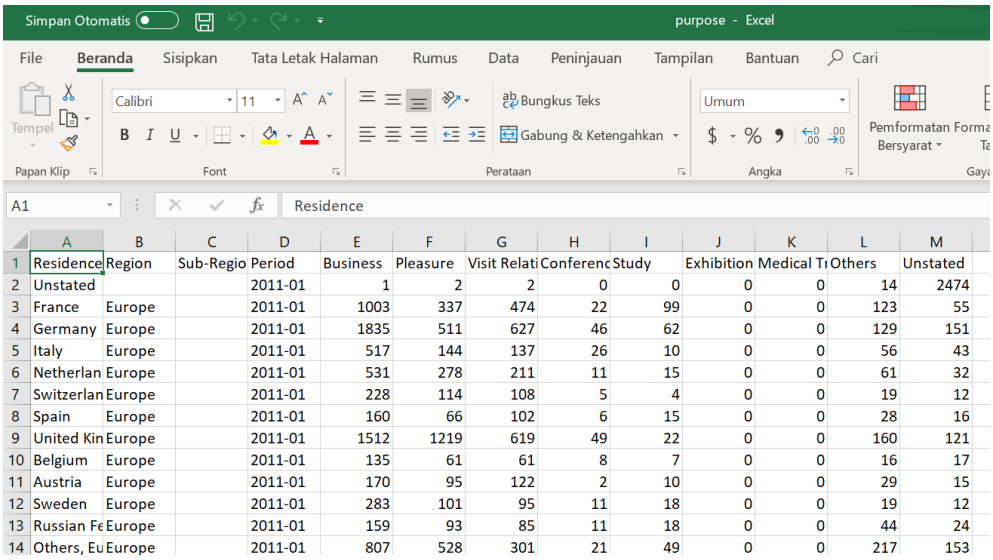
Model Evaluation is an integral part of the model development process. It helps to find the best model that represents our data and how well the chosen model will work in the future. Evaluating model performance with the data used for training is not acceptable in data mining because it can easily generate overoptimistic and overfitted models. There are two methods of evaluating models in data mining, Hold-Out and Cross-Validation. To avoid overfitting, both methods use a test set (not seen by the model) to evaluate model performance. In this work, the default classification will be used using Bayes Naïve algorithm. Figure 12 shows the result after preprocessing data been done.



Picture 15. Evaluation

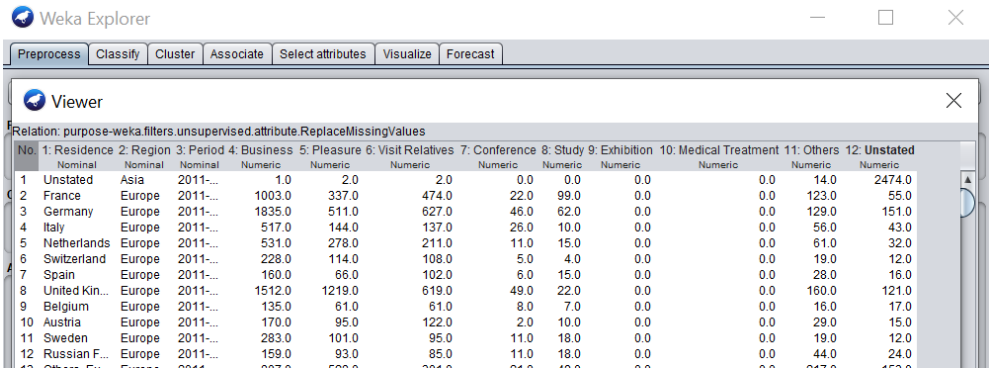
4. Result and Presentation

A. Example of Row Data



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Residence	Region	Sub-Regio	Period	Business	Pleasure	Visit Relati	Conferenc	Study	Exhibition	Medical Tr	Others	Unstated
2	Unstated			2011-01	1	2	2	0	0	0	0	14	2474
3	France	Europe		2011-01	1003	337	474	22	99	0	0	123	55
4	Germany	Europe		2011-01	1835	511	627	46	62	0	0	129	151
5	Italy	Europe		2011-01	517	144	137	26	10	0	0	56	43
6	Netherlan	Europe		2011-01	531	278	211	11	15	0	0	61	32
7	Switzerlan	Europe		2011-01	228	114	108	5	4	0	0	19	12
8	Spain	Europe		2011-01	160	66	102	6	15	0	0	28	16
9	United Kin	Europe		2011-01	1512	1219	619	49	22	0	0	160	121
10	Belgium	Europe		2011-01	135	61	61	8	7	0	0	16	17
11	Austria	Europe		2011-01	170	95	122	2	10	0	0	29	15
12	Sweden	Europe		2011-01	283	101	95	11	18	0	0	19	12
13	Russian Fe	Europe		2011-01	159	93	85	11	18	0	0	44	24
14	Others, Eu	Europe		2011-01	807	528	301	21	49	0	0	217	153

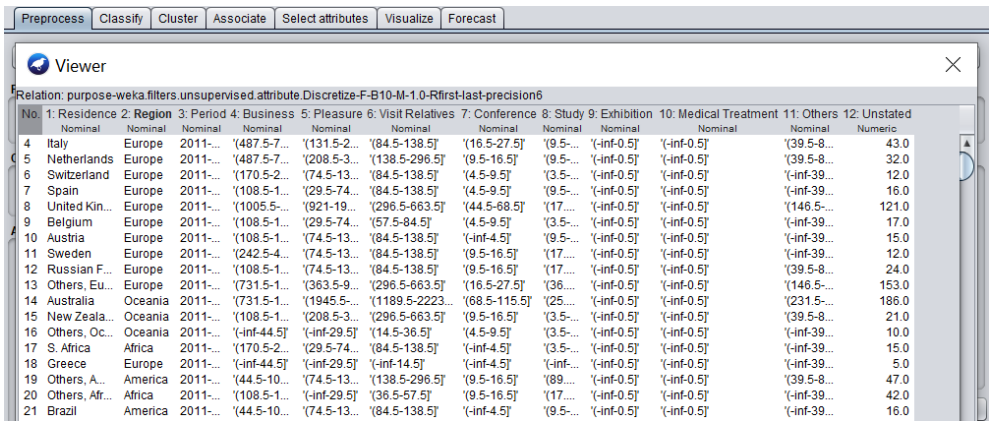
Picture 16. Example of raw Data



Relation: purpose-weka.filters.unsupervised.attribute.ReplaceMissingValues

No.	1: Residence Nominal	2: Region Nominal	3: Period Nominal	4: Business Numeric	5: Pleasure Numeric	6: Visit Relatives Numeric	7: Conference Numeric	8: Study Numeric	9: Exhibition Numeric	10: Medical Treatment Numeric	11: Others Numeric	12: Unstated Numeric
1	Unstated	Asia	2011-...	1.0	2.0	2.0	0.0	0.0	0.0	0.0	14.0	2474.0
2	France	Europe	2011-...	1003.0	337.0	474.0	22.0	99.0	0.0	0.0	123.0	55.0
3	Germany	Europe	2011-...	1835.0	511.0	627.0	46.0	62.0	0.0	0.0	129.0	151.0
4	Italy	Europe	2011-...	517.0	144.0	137.0	26.0	10.0	0.0	0.0	56.0	43.0
5	Netherlands	Europe	2011-...	531.0	278.0	211.0	11.0	15.0	0.0	0.0	61.0	32.0
6	Switzerland	Europe	2011-...	228.0	114.0	108.0	5.0	4.0	0.0	0.0	19.0	12.0
7	Spain	Europe	2011-...	160.0	66.0	102.0	6.0	15.0	0.0	0.0	28.0	16.0
8	United Kin...	Europe	2011-...	1512.0	1219.0	619.0	49.0	22.0	0.0	0.0	160.0	121.0
9	Belgium	Europe	2011-...	135.0	61.0	61.0	8.0	7.0	0.0	0.0	16.0	17.0
10	Austria	Europe	2011-...	170.0	95.0	122.0	2.0	10.0	0.0	0.0	29.0	15.0
11	Sweden	Europe	2011-...	283.0	101.0	95.0	11.0	18.0	0.0	0.0	19.0	12.0
12	Russian F...	Europe	2011-...	159.0	93.0	85.0	11.0	18.0	0.0	0.0	44.0	24.0

Picture 17. Exampled of cleande Data



Relation: purpose-weka.filters.unsupervised.attribute.Discretize-F-B10-M-1.0-Rfirst-last-precision6

No.	1: Residence Nominal	2: Region Nominal	3: Period Nominal	4: Business Nominal	5: Pleasure Nominal	6: Visit Relatives Nominal	7: Conference Nominal	8: Study Nominal	9: Exhibition Nominal	10: Medical Treatment Nominal	11: Others Nominal	12: Unstated Numeric
4	Italy	Europe	2011-...	'(487.5-7...	'(131.5-2...	'(84.5-138.5]	'(16.5-27.5]	'(9.5-...	'(-inf-0.5]	'(-inf-0.5]	'(39.5-8...	43.0
5	Netherlands	Europe	2011-...	'(487.5-7...	'(208.5-3...	'(138.5-296.5]	'(9.5-16.5]	'(9.5-...	'(-inf-0.5]	'(-inf-0.5]	'(39.5-8...	32.0
6	Switzerland	Europe	2011-...	'(170.5-2...	'(74.5-13...	'(84.5-138.5]	'(4.5-9.5]	'(3.5-...	'(-inf-0.5]	'(-inf-0.5]	'(-inf-39...	12.0
7	Spain	Europe	2011-...	'(108.5-1...	'(29.5-74...	'(84.5-138.5]	'(4.5-9.5]	'(9.5-...	'(-inf-0.5]	'(-inf-0.5]	'(-inf-39...	16.0
8	United Kin...	Europe	2011-...	'(1005.5-...	'(921-19...	'(296.5-663.5]	'(44.5-68.5]	'(17....	'(-inf-0.5]	'(-inf-0.5]	'(146.5-...	121.0
9	Belgium	Europe	2011-...	'(108.5-1...	'(29.5-74...	'(57.5-84.5]	'(4.5-9.5]	'(3.5-...	'(-inf-0.5]	'(-inf-0.5]	'(-inf-39...	17.0
10	Austria	Europe	2011-...	'(108.5-1...	'(74.5-13...	'(84.5-138.5]	'(-inf-4.5]	'(9.5-...	'(-inf-0.5]	'(-inf-0.5]	'(-inf-39...	15.0
11	Sweden	Europe	2011-...	'(242.5-4...	'(74.5-13...	'(84.5-138.5]	'(9.5-16.5]	'(17....	'(-inf-0.5]	'(-inf-0.5]	'(-inf-39...	12.0
12	Russian F...	Europe	2011-...	'(108.5-1...	'(74.5-13...	'(84.5-138.5]	'(9.5-16.5]	'(17....	'(-inf-0.5]	'(-inf-0.5]	'(39.5-8...	24.0
13	Others, Eu...	Europe	2011-...	'(731.5-1...	'(363.5-9...	'(296.5-663.5]	'(16.5-27.5]	'(36....	'(-inf-0.5]	'(-inf-0.5]	'(146.5-...	153.0
14	Australia	Oceania	2011-...	'(731.5-1...	'(1945.5-...	'(1189.5-2223...	'(68.5-115.5]	'(25....	'(-inf-0.5]	'(-inf-0.5]	'(231.5-...	186.0
15	New Zeala...	Oceania	2011-...	'(108.5-1...	'(208.5-3...	'(296.5-663.5]	'(9.5-16.5]	'(3.5-...	'(-inf-0.5]	'(-inf-0.5]	'(39.5-8...	21.0
16	Others, Oc...	Oceania	2011-...	'(-inf-44.5]	'(-inf-29.5]	'(14.5-36.5]	'(4.5-9.5]	'(3.5-...	'(-inf-0.5]	'(-inf-0.5]	'(-inf-39...	10.0
17	S. Africa	Africa	2011-...	'(170.5-2...	'(29.5-74...	'(84.5-138.5]	'(-inf-4.5]	'(3.5-...	'(-inf-0.5]	'(-inf-0.5]	'(-inf-39...	15.0
18	Greece	Europe	2011-...	'(-inf-44.5]	'(-inf-29.5]	'(-inf-14.5]	'(-inf-4.5]	'(-inf-...	'(-inf-0.5]	'(-inf-0.5]	'(-inf-39...	5.0
19	Others, A...	America	2011-...	'(44.5-10...	'(74.5-13...	'(138.5-296.5]	'(9.5-16.5]	'(89....	'(-inf-0.5]	'(-inf-0.5]	'(39.5-8...	47.0
20	Others, Afr...	Africa	2011-...	'(108.5-1...	'(-inf-29.5]	'(36.5-57.5]	'(9.5-16.5]	'(17....	'(-inf-0.5]	'(-inf-0.5]	'(-inf-39...	42.0
21	Brazil	America	2011-...	'(44.5-10...	'(74.5-13...	'(84.5-138.5]	'(-inf-4.5]	'(9.5-...	'(-inf-0.5]	'(-inf-0.5]	'(-inf-39...	16.0

Picture 18. Example of discratized Data

5. Conclusion

In this report, it shows how the pre-processing data for data cleaning and reduction process applied to the Tourism Bureau, MOTC Republic of China dataset. The extraction of the data monthly number of visitors to Taiwan by their residence and purpose. per month become easy after the pre-processing has been done. Through this step we can select the best features contribute to this task before applying classification process

Acknowledgement

I would first like to thank my lecture Prof Dr Azuraliza Abu Bakar of the Faculty of Information Science and Technology Universiti Kebangsaan Malaysia. For teach ang guide me in learning the fundamental of data science.

Finally, I must express my very profound gratitude to my parents and to my wife Mutia Perdana Cifa for providing me with unfailing support and continuous encouragement my study.This accomplishment would not have been possible without them. Thank you

Referensi

[1] Chen, C. C., & Lin, Y. H. (2012). Segmenting mainland Chinese tourists to Taiwan by destination familiarity: A factor - cluster approach. International Journal of Tourism Research, 14(4), 339-352.

[2] Hadjikakou, M., Chenoweth, J., Miller, G., Druckman, A., & Li, G. (2014). Rethinking the economic contribution of tourism: case study from a Mediterranean Island. Journal of Travel Research, 53(5), 610-624.

[3] Dev, V., Tyagi, A., & Singh, P. (2017). Tourism Demand Forecasting and Management. International Journal of Business and Management Invention, ISSN (Online): 2319–8028, ISSN (Print): 2319–801X, 6(2), 01-09.

[4] Elena, M., Lee, M. H., Suhartono, H., Hossein, I., Rahman, N. H. A., & Bazilah, N. A. (2012). Fuzzy time series and sarima model for forecasting tourist arrivals to bali. Jurnal Teknologi, 57(1).