

data exploration and preparation

Assignment 2



**Kunanon Pititheerachot**

**12634123**

Table of Contents

[Initial data exploration 2](#_Toc513240074)

[Attributes 2](#_Toc513240075)

[row\_ID 2](#_Toc513240076)

[add\_these\_pw\_job\_title\_9089 2](#_Toc513240077)

[agent\_city 2](#_Toc513240078)

[agent\_firm\_name 4](#_Toc513240079)

[agent\_state 5](#_Toc513240080)

[application\_type 6](#_Toc513240081)

[case\_no 6](#_Toc513240082)

[case\_number 6](#_Toc513240083)

[case\_received\_date 6](#_Toc513240084)

[case\_status 7](#_Toc513240085)

[class\_of\_admission 8](#_Toc513240086)

[country\_of\_citizenship 9](#_Toc513240087)

[country\_of\_citzenship 9](#_Toc513240088)

[decision\_date 10](#_Toc513240089)

[employer\_address\_1 11](#_Toc513240090)

[employer\_address\_2 12](#_Toc513240091)

[employer\_city 12](#_Toc513240092)

[employer\_country 13](#_Toc513240093)

[employer\_decl\_info\_title 13](#_Toc513240094)

[employer\_name 14](#_Toc513240095)

[employer\_num\_employees 15](#_Toc513240096)

[employer\_phone 17](#_Toc513240097)

[employer\_phone\_ext 18](#_Toc513240098)

[employer\_postal\_code 18](#_Toc513240099)

[employer\_state 19](#_Toc513240100)

[employer\_yr\_estab 20](#_Toc513240101)

[foreign\_worker\_info\_alt\_edu\_experience 21](#_Toc513240102)

[foreign\_worker\_info\_birth\_country 21](#_Toc513240103)

[foreign\_worker\_info\_city 22](#_Toc513240104)

[foreign\_worker\_info\_education 23](#_Toc513240105)

[Clusters and Outliers 25](#_Toc513240106)

[Employer Number Employees and Employer Year Establish 25](#_Toc513240107)

[Employer Year Establish and Employer State 26](#_Toc513240108)

[Employer number of employees and state 27](#_Toc513240109)

[Interesting Finding in Attributes 28](#_Toc513240110)

[Data pre-processing 28](#_Toc513240111)

[Binning 28](#_Toc513240112)

[Normalisation 31](#_Toc513240113)

[Discretisation 32](#_Toc513240114)

[Binarisation 33](#_Toc513240115)

[Summary 34](#_Toc513240116)

[Case received date 34](#_Toc513240117)

[Decision date 34](#_Toc513240118)

[Employer state 35](#_Toc513240119)

[Employer year establish 35](#_Toc513240120)

[Employer number employee 35](#_Toc513240121)

[Clusters and Outliers 35](#_Toc513240122)

[Employer Number Employees and Employer Year Establish 35](#_Toc513240123)

[Employer Year Establish and Employer State 35](#_Toc513240124)

[Conclusion 35](#_Toc513240125)

# Initial data exploration

## Attributes

row\_ID

‘row ID’ is changed to ‘row\_ID’ in the data cleaning process in order to make the same format of the other columns’ name.

#### Attribute type:

row\_ID column is a nominal attribute because it has differently numbers in each row that not represent the relationship between each other and they are a unique name for data records.

#### Range:

The row\_ID attribute, it’s range start from 94 to 374257

add\_these\_pw\_job\_title\_9089

Attribute ‘add\_these\_pw\_job\_title\_9089’ is an empty column

agent\_city

In the data cleaning process in order to make the same format of the data such as ALBANY -> Albany, ALPHARETTA -> Alpharetta and etc., inside csv file before import to tool called Jupyter notebook.

#### Attribute type:

‘agent\_city’ is a nominal attribute as it cannot be ordered, the data value in the column represent the name of the city.

#### Value Range:

There are 254 different cities name in this dataset that mostly cities are in United State, and the top is ‘None’ that data are replaced for the empty data and it is 242 data out of 2000(Figure1).

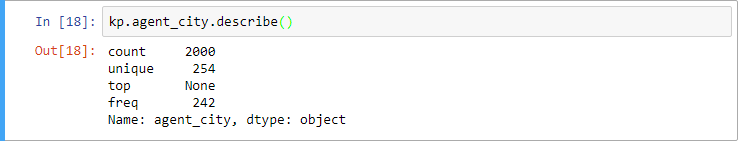


Figure 1: description of agent\_city.

#### Frequency & Distribution:

As I mentioned before this data set is big as it has 254 different cities name which is cause an inconveniently to represent all of them in this one bar graph. Figure 2 illustrate the distribution and frequency of agent cities data set which represent the most frequent data is ‘None’ with 254 occurrences as mentioned above, followed by New York with 178 occurrences, and San Francisco with 173 occurrences.

Figure 2: Distribution and Frequently of agent cities.

### agent\_firm\_name

In the data cleaning process in order to make the count of empty data set I replace ‘None’ to all empty data set before plot graph and counting data.

#### Attribute type:

‘agent\_firm\_name’ is a nominal attribute as it cannot be ordered, the data value in the column represent the firm name.

#### Value Range:

There are 696 different firm names in this dataset, the top occurrences data is ‘None’ with 276 data out of 2000 which is represent in ‘Figure 3’ below, if except for ‘None’ the top of the data should be firm name called ‘Fragomen, Del Rey, Bernsen & Loewy, LLP’ with 238 occurrences.

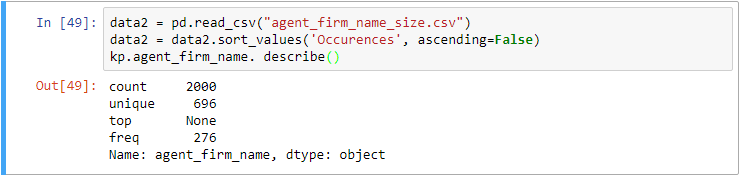


Figure 3: Description of agent firm name.

#### Frequency & Distribution:

As it has 696 different firm name which is cause an inconveniently to represent all of them. Figure 4 illustrate the distribution and frequency of agent firm name data set which represent the most frequent data is ‘None’ as mentioned above, followed by ‘Fragomen, Del Rey, Bernsen & Loewy, LLP’ with 238 occurrences, then ‘Berry Appleman & Leiden LLP0’ with 130 occurrences, and ‘Fragomen, Del Rey, Bernsen & Loewy LLP’ with 31 occurrences.

Figure 4: Distribution and frequently of agent firm name.

### agent\_state

In the data cleaning process in order to make the count of empty data set I replace ‘None’ to all empty data set before plot graph and counting data. In addition, some of data set come with different type such as NEW YORK and NY so I made some change to all of the abbreviation version to full of state name such as AL -> ALABAMA, and TX -> TEXAS, and so on.

#### Attribute type:

‘agent \_state’ is a nominal attribute as it cannot be ordered, the data value in the column represent the state name and it not be more valuable for which state has come before.

#### Value Range:

There are 39 different firm names in this dataset, the top occurrences data is ‘CALIFORNIA’ with 458 data out of 2000 which is represent in ‘Figure 5’ below.



Figure 5: Description of agent state

#### Frequency & Distribution:

Figure 6 illustrate the distribution and frequency of 39 different agent state data set which represent the most frequent data is ‘CALIFORNIA’ with 458 occurrences as mentioned above, followed by ‘None’ with 268 occurrences, then ‘NEW YORK’ with 199 occurrences, and ‘TEXAS’ with 156 occurrences.

Figure 6: Distribution and frequently of agent state.

### application\_type

Attribute ‘application\_type’is an empty column

### case\_no

Attribute ‘case\_no’ is an empty column

### case\_number

#### Attribute type:

For case\_number column is a nominal attribute because it has differently numbers in each row that not represent the relationship between each other and they are just a uniquely name for data records.

#### Range:

The row\_ID attribute, it’s range can be ordering as ascending from A-06138-19016 to A-16351-81653

### case\_received\_date

#### Attribute type:

For case\_received\_date is considered as ordinal attribute. For instant, case received date on 17/03/2009 must be occurred before case received on 16/12/2016 and year 2016 case received must have passed year 2015 and before.

#### Value Range:

There are 794 different case received date in this dataset which is range start from 17/03/2009 to 16/12/2016, the top occurrences data is on ‘26/04/2016’ with 9 occurrences which is represent in ‘Figure 7’.



Figure 7: Description of case received date.

#### Frequency & Distribution:

Figure 8 illustrate about distribution and frequency of case received date, the top occurrence is 9 on 26/04/2016. Moreover, in figure 9 represent more details about in each year case received date is different as pie chart represent the year 2009 to 2012 is come with 0% of total occurrences but since 2013 trend has slightly increase to 7% and dramatically increase to 27% in 2014, the occurrences continue rise to the peak in 2015 with 36% and slightly drop in 2016 with 30% of occurrences.

Figure 8: Distribution and frequently of case received date.

Figure 9: Pie graph illustrate distribution and frequently of case received date.

### case\_status

#### Attribute type:

‘case\_status’ is a nominal attribute as it cannot be ordered as it does not have quantitative value, the data value in the column represent the status of cases and it not be more valuable for which status has come before.

#### Value Range:

There are 4 different types of case status which are ‘Certified’, ‘Certified-Expired’, ‘Denied’, and Withdrawn the top occurrences data is ‘Certified’ with 1016 concurrences out of 2000 which is represent in ‘Figure 10’.

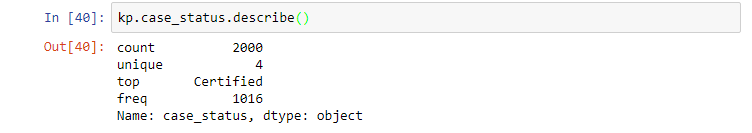


Figure 10: Description of case status.

#### Frequency & Distribution:

Figure 11 illustrate the distribution and frequency of 4 different case status, the most frequent data is ‘Certified’ as mentioned above, followed by ‘Certified-Expired’ with 819 occurrences, then ‘Withdrawn’ with 86 occurrences, and ‘Denied’ with 79 occurrences.

Figure 11: Distribution and frequently of case status.

### class\_of\_admission

In the data cleaning process in order to make the count of empty data set I replace ‘None’ to all empty data set before plot graph and counting data.

#### Attribute type:

‘class\_of\_admission’ is a nominal attribute as it cannot be ordered, this data set is provided unclear meaning with one English letter and number which meaning is unknown assume that it all act as labels.

### country\_of\_citizenship

#### Attribute type:

‘country\_of\_citizenship’ is a nominal attribute as it cannot be ordered, the data value in the column represent the country name and it not be more valuable for which country has come before.

#### Value Range:

There are 81 different country names in this dataset, the top occurrences data is ‘INDEA’ with 1205 data out of 2000 which is represent in ‘Figure 12’.

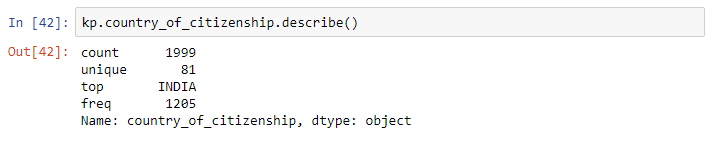
****

Figure 12: Description of country of citizenship.

#### Frequency & Distribution:

Figure 13 illustrate the distribution and frequency of 81 different country of citizenship, the most frequent data is ‘INDIA’ as mentioned above, followed by ‘CHINA’ with 173 occurrences, ‘SOUTH KOREA’ with 100 occurrences, and ‘CANADA’ with 67 occurrences. And this graph is representing with only partial of the data set as 81 countries cannot fit properly in one histogram graph.

Figure 13: Distribution and frequently of country of citizenship.

### country\_of\_citzenship

Attribute ‘country\_of\_citizinship’ is an empty column

### decision\_date

#### Attribute type:

For decision\_date is considered as ordinal attribute. For instant, case received date on 01/10/2014 must be occurred before case received on 30/12/2016 and year 2016 decision date must have passed year 2015 and before.

#### Value Range:

There are 534 different decision dates in this dataset which is range start from 01/10/2014 to 30/12/2016(Figure 14), the top occurrences data is on ‘19/11/2015’ and ‘30/07/2015’ with 15 occurrences which is represent in ‘Figure 15’.

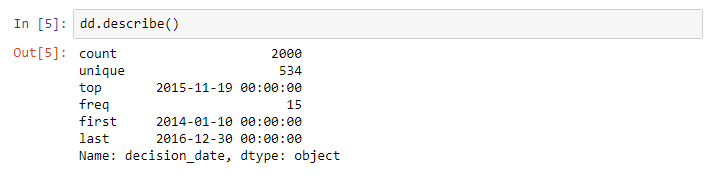


Figure 14: Description of decision date.

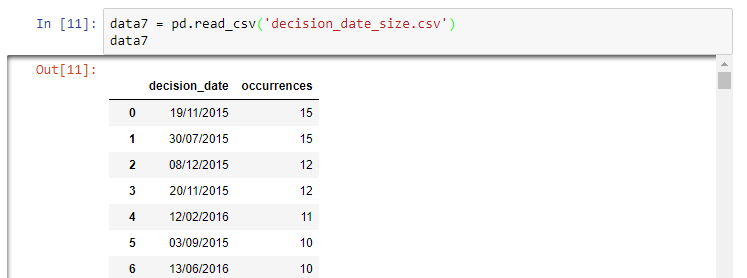


Figure 15: Description of decision date.

#### Frequency & Distribution:

Figure 16 illustrate about distribution and frequency of decision date, the top occurrence is 15 on 19/11/2015 and 30/07/2015, with 534 different decision date it is inconvenient to include all data set in one histogram graph. Moreover, in figure 17 represent more details about in each year decision date is different as pie chart represent the year 2014 to 2016 is come with mostly in 2016 with 51%(1019) of total occurrences, the second biggest is in 2015 with 41%(827) and 8% in 2014 with 154 occurrences.

Figure 16: Distribution and frequently of case status.

Figure 17: Pie graph illustrate distribution and frequently of decision date.

### employer\_address\_1

‘employee\_address\_1’ is changed to ‘employer\_address\_1’ in the data cleaning process in order to make the same format to proper letter.

#### Attribute type:

Employee\_address\_1 data set is a nominal attribute because it has differently numbers in each row that not represent the relationship between each other and they are a unique name for data records.

### employer\_address\_2

‘employee\_address\_2’ is changed to ‘employer\_address\_2’ in the data cleaning process in order to make the same format to proper letter.

#### Attribute type:

Employee\_address\_2 data set is a nominal attribute because it has some duplicate of numbers in each row but that not represent the relationship between each other and they are a uniquely name act as label for data records.

### employer\_city

‘employee\_city’ is changed to ‘employer\_city’ in the data cleaning process in order to make the same format to proper letter such as Kenansville -> KENANSVILLE.

#### Attribute type:

‘employer\_city’ is a nominal attribute as it cannot be ordered, the data value in the column represent the name of the city.

#### Value Range:

There are 519 different cities name in this dataset in United State of America, and the top is ‘COLLEGE STATION’ that data are replaced for the empty data and it is 126 data out of 2000(Figure18).

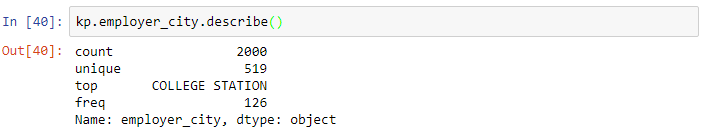


Figure 18: Description of employer city.

#### Frequency & Distribution:

Figure 19 illustrate the distribution and frequency of 519 different cities of United State of America, the most frequent data is ‘COLLEGE STATION’ as mentioned above, followed by ‘NEW YORK’ with 84 occurrences, ‘MOUNTAIN VIEW’ with 57 occurrences, and ‘SANTA CLARA’ with 54 occurrences. And this graph is representing with only partial of the data set as 519 cities cannot fit properly in one histogram graph.

Figure 19: Description and frequently of employer city.

### employer\_country

#### Attribute type:

‘employer\_city’ is a nominal attribute as it cannot be ordered, the data represent the exactly same value which is ‘UNITED STATES OF AMERICA’

### 

### employer\_decl\_info\_title

In the data cleaning process in order to make the same format to proper letter such as DIRECTOR OF HUMAN RESOURCES -> Director of Human Resources.

#### Attribute type:

‘employer\_decl\_info\_title is a nominal attribute as it cannot be ordered because it has no quantitative order which make it has no hierarchy, the data value in the column represent the name of employer title, for instant this data set represent of Chief of Staff and Hr Manager that which one is better but only act as labels.

#### Value Range:

There are 649 different title names in this dataset, the top occurrences data is ‘President’ with 233 data out of 2000 which is represent in ‘Figure 20’ below.

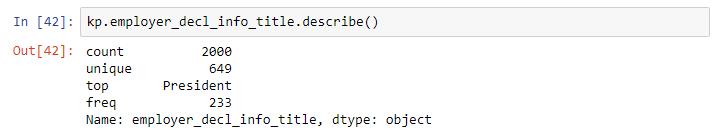


Figure 20: Description of employer decl info title.

#### Frequency & Distribution:

Figure 21 illustrate the distribution and frequency of 649 different title names data set which represent the most frequent data is ‘President’ with 233 occurrences as mentioned above, followed by ‘Immigration Specialist’ with 194 occurrences, and ‘Ceo’ with 53 occurrences.

Figure 21: Distribution and frequently of employer decl info title

### employer\_name

#### Attribute type:

‘employer\_name’ is a nominal attribute as it cannot be ordered, the data value in the column represent the employer company name.

#### Value Range:

There are 1277 different firm names in this dataset, the top occurrences data is ‘COGNIZANT TECHNOLOGY SOLUTIONS US COPORATION’ with 126 data out of 2000 which is represent in ‘Figure 22’ below.

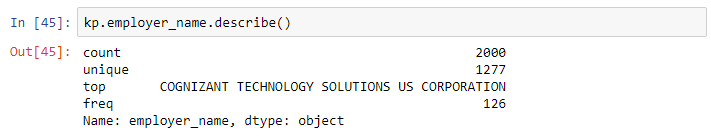


Figure 22: Description of employer name.

#### Frequency & Distribution:

As it has 1277 different employer name which is cause an inconveniently to represent all of them. Figure 23 illustrate the distribution and frequency of employer name data set which represent the second most frequent data is ‘MICROSOFT CORPORATION’ with 39 occurrences, then ‘GOOGLE INC.’ with 37 occurrences, and ‘AMAZON CORPORATE LLC’ with 33 occurrences.

Figure 23: Distribution and frequently of employer name.

### employer\_num\_employees

In the data cleaning process one row of them has empty data so I replace with number 0.

#### Attribute type:

‘employer\_num\_employees’ is a ratio attributes because it has non-negative number and can be ordered by using numeric values.

#### Statistic:

|  |  |
| --- | --- |
| **Measurements** | **Value** |
| Range | 0 (min) to 2200000 (max) |
| Median | 1600 |
| Mode | Mode = array([29000], count = array([52]) |
| Average | 24565.856 |
| Variance | 8215686921.713119 |
| Standard Deviation | 90640.42653095316 |
| 25ᵀᴴ Percentile | 85.0 |
| 50ᵀᴴ Percentile | 1600.0 |
| 75ᵀᴴ Percentile | 22000.00 |
| 99ᵀᴴ Percentile | 256420.0 |

Table 1: Table of employer num employees.

#### Frequency & Distribution:

As it has a lot of different ‘employer num employees’ data set as shown in figure 25 which is cause an inconveniently to represent all of them in one histogram graph. Figure 24 illustrate the distribution and frequency of ‘employer num employees’ data set which represent the most frequent data is employer with 29000 employees with 52 occurrences, then 35000 employees with 47 occurrences, and 60000 with 39 occurrences.

Figure 24: Distribution and frequently of employer num employees.



Figure 25: Box graph illustrate Distribution and frequently of employer num employees.

### employer\_phone

In the data cleaning process in order to make the same format to proper phone ‘s format such as 9102961521 -> (910) 996-1521.

#### Attribute type:

Employer\_phone data set is a nominal attribute because it has differently numbers in each row that not represent the relationship between each other and they are a unique phone number for data records.

#### Value Range:

There are 1264 different phone number in this dataset, the top occurrences data is ‘(201) 290-9573’ with 126 data out of 2000 which is represent in ‘Figure 26’ below.

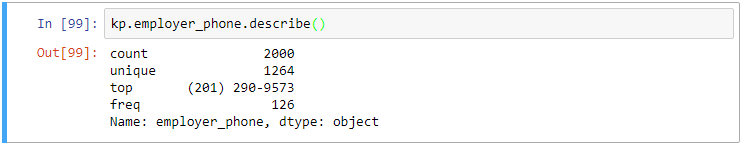


Figure 26: Description of employer phone.

#### Frequency & Distribution:

As it has a lot of unique data set as shown in data set which is cause an inconveniently to represent all of them in one histogram graph. Figure 27 illustrate the distribution and frequency of ‘employer phone’ data set which represent the second most frequent data is phone number ‘(206) 266-1000’ with 39 occurrences, then ‘(650) 253-0000’ with 37 occurrences.

Figure 27: Distribution and frequently of employer phone.

### employer\_phone\_ext

#### Attribute type:

Employer\_phone data set is a nominal attribute because it has differently numbers and most of rows are empty and not represent the relationship between each other and they are an extension of phone number for data records.

### employer\_postal\_code

#### Attribute type:

Employer\_postal\_code data set is a nominal attribute because it has differently numbers that not represent the relationship between each other and they are just post code number data records.

#### Value Range:

There are 825 different employer postal code in this dataset, the top occurrences data is ‘77845’ with 126 data out of 2000 which is represent.

#### Frequency & Distribution:

As it has a lot of different ‘employer postal code’ data set as shown which is cause an inconveniently to represent all of them in one histogram graph. Figure 28 illustrate the distribution and frequency of ‘employer postal code’ data set which represent the second most frequent data is ‘94043’ with 55 occurrences, then ‘98052’ with 45 occurrences.

Figure 28: Distribution and frequently of employer postal code.

### employer\_state

In the data cleaning process some of data set come with different type such as NEW YORK and NY so I made some change to all of the abbreviation version to full of state name such as AL -> ALABAMA, and TX -> TEXAS, and so on.

#### Attribute type:

‘employer\_state’ is a nominal attribute as it cannot be ordered, the data value in the column represent the state name and it not be more valuable for which state has come before.

#### Value Range:

There are 50 different firm names in this dataset, the top occurrences data is ‘CALIFORNIA’ with 536 data out of 2000 which is represent.

#### Frequency & Distribution:

Figure 29 illustrate the distribution and frequency of 50 different employer state data set which represent the most frequent data is ‘CALIFORNIA’ with 536 occurrences as mentioned above, followed by ‘TEXAS’ with 290 occurrences, then ‘NEW JERSEY’ with 160 occurrences, and ‘NEW YORK’ with 123 occurrences.

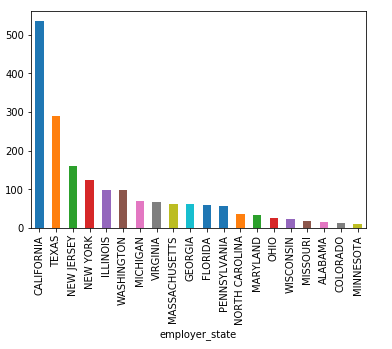


Figure 29: Distribution and frequently of employer state.

### employer\_yr\_estab

#### Attribute type:

‘employer\_yr\_estab’ is a ratio attributes because it has non-negative number and can be ordered by using numeric values. For example, the year in data set start 1740 that has passed before the end of data set which is 2016 which means gap value is 276 years.

#### Statistic:

|  |  |
| --- | --- |
| **Measurements** | **Value** |
| Range | 1740 (min) to 2016 (max) |
| Median | 1997 |
| Mode | mode=array([1994], dtype=int64), count=array([161]) |
| Average | 1985.2295 |
| Variance | 1264.6561578289145 |
| Standard Deviation | 35.562004412419085 |
| 25ᵀᴴ Percentile | 1982.0 |
| 50ᵀᴴ Percentile | 1997.0 |
| 75ᵀᴴ Percentile | 2004.0 |
| 99ᵀᴴ Percentile | 2014.0 |

Table 2: Table of employer num employees.

#### Frequency & Distribution:

Figure 30 illustrate the distribution and frequency of ‘employer yr estab’ data set which represent the most frequent data is year ‘1994’ with 161 occurrences, the second is ‘1999’ with 110 occurrences, and ‘1998’ with 106 occurrences.

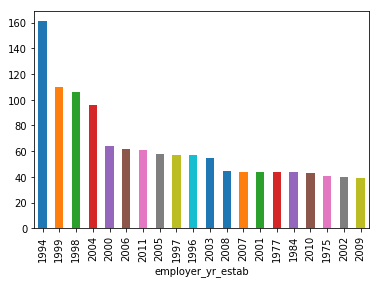


Figure 30: Distribution and frequently of employer yr estab.

### foreign\_worker\_info\_alt\_edu\_experience

#### Attribute type:

foreign\_worker\_info\_alt\_edu\_experience data set is a nominal attribute because it has differently one letter and most of rows are empty and not represent the relationship between each other.

### foreign\_worker\_info\_birth\_country

#### Attribute type:

foreign\_worker\_info\_birth\_country data set is a nominal attribute because it has differently name of countries and most of rows are empty and not represent the relationship between each other.

#### Value Range:

This data set has 60 different country names.

#### Frequency & Distribution:

Figure 31 illustrate the distribution and frequency of partial of different employee birth countries data set which represent the most frequent data is ‘INDIA’ with 469 occurrences, followed by ‘CHINA’ with 67 occurrences, then ‘SOUTH KOREA’ with 49 occurrences, and ‘MEXICO’ with 14 occurrences.

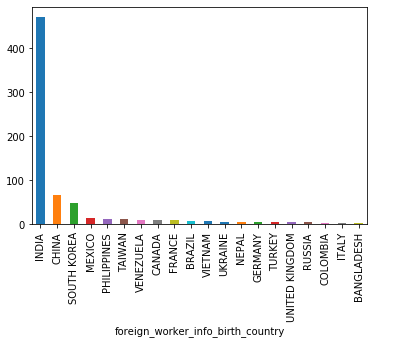


Figure 31: Distribution and frequently of foreign worker info birth country.

### foreign\_worker\_info\_city

#### Attribute type:

foreign\_worker\_info\_city data set is a nominal attribute because it has differently name of countries and most of rows are empty and not represent the relationship between each other only act as labels.

#### Value Range:

This data set has 720 different city names.

#### Frequency & Distribution:

Figure 32 illustrate the distribution and frequency of partial of different employee info city data set which represent the most frequent data is ‘SUNNYVALE’ with 62 occurrences, followed by ‘SAN JOSE’ with 61 occurrences, then ‘FREMONT’ with 43 occurrences, and ‘JERSEY CITY’ with 38 occurrences.

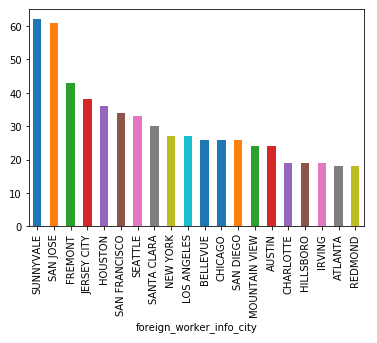


Figure 32: Distribution and frequently of foreign worker info city.

### foreign\_worker\_info\_education

#### Attribute type:

foreign\_worker\_info\_education data set is an ordinal attribute because it has hierarchy in education level. For instant, before taking the Doctorate degree student must have graduated from Master’s degree and before that is Bachelor’s degree.

#### Value Range:

This data set has 5 different education levels which are ‘Hight School’, ‘Associate’s’, ‘Bachelor’s’, ‘Master’s’ and ‘Doctorate’.

#### Frequency & Distribution:

Figure 33 illustrate the distribution and frequently of ‘foreign worker info education’ set which represent the most frequent data is ‘Master’s’ with 960 occurrences, followed by ‘Bachelor’s’ with 847 occurrences, then ‘Doctorate’ with 119 occurrences, ‘High School’ with 50 occurrences, and ‘Associate’s’ with 24 occurrences. And figure 33 represent more details about education level with box graph.

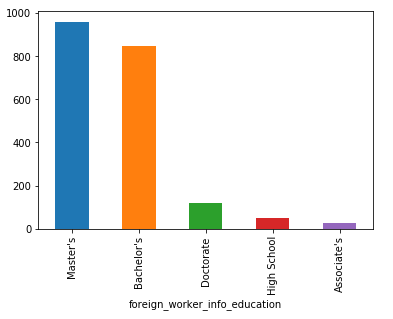


Figure 32: Distribution and frequently of foreign worker info education.

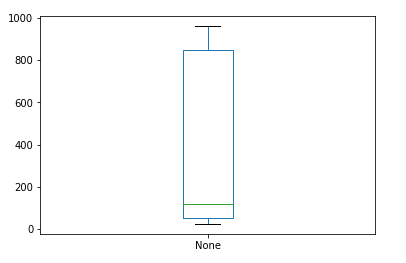


Figure 33: Box graph illustrate distribution and frequently of foreign worker info education.

## Clusters and Outliers

The tools that were used in this section are several but the mainly used environment is Jupyter notebook that use Python programming language for data implementation and the several libraries were used such as Pandas, Numpy, Scipy, and Matplotlib.pyplot in order to visualise and summarise the data set, and mainly function that has used in order to find interesting clusters and outliers of data is scatter plot and some box plot.

### Employer Number Employees and Employer Year Establish

Figure 34 illustrate the scatter plot graph made in Jupyter notebook using two set of selected randomly data from ‘employer num employees’ and ‘employer yr estab’. ‘employer num employees’ represents the number of employees for individual employer in Y-axis. ‘employer yr estab’ represents the establish year for individual employer in X-axis. This graph represents the relationship between number of employees and year establish, the main cluster has shown in green square which represent that mostly employers have less than 50,000 employees and significant year are around year establish in 2000. On the other hand, red squares represent the outliers’ data which are less values which one of the data represent that one employer has more than 2,000,000 employees and one of data has nearly 0 in year establish 1800.



Figure 34: Scatter plot of employee number and employer year establish

### Employer Year Establish and Employer State

In next couple of selected randomly data sets are ‘employer yr estab’ and ‘employer state’ which are represented connection between represent each employer year of establish by individual year in X-axis and employer state represent state by individuals name in the United State Country in Y-axis. As scatter plot shown below (Figure 35) that mostly of data which is considered as main cluster is within year establish after year 1950 to after 2000, and including several states such as Illinois, Michigan, Arizona, Georgia, California, Texas, and New Jersey. In addition, the most common employer state is in California in year around 2000. For the outliers in this graph represent in red squares which identified as less value data such as around year 1800 only one employer ‘New Jersey’ and in year around 2000 they represent only 1 employer in Kansas state which these outliers could lead us to the unnecessary meaning for this data set.

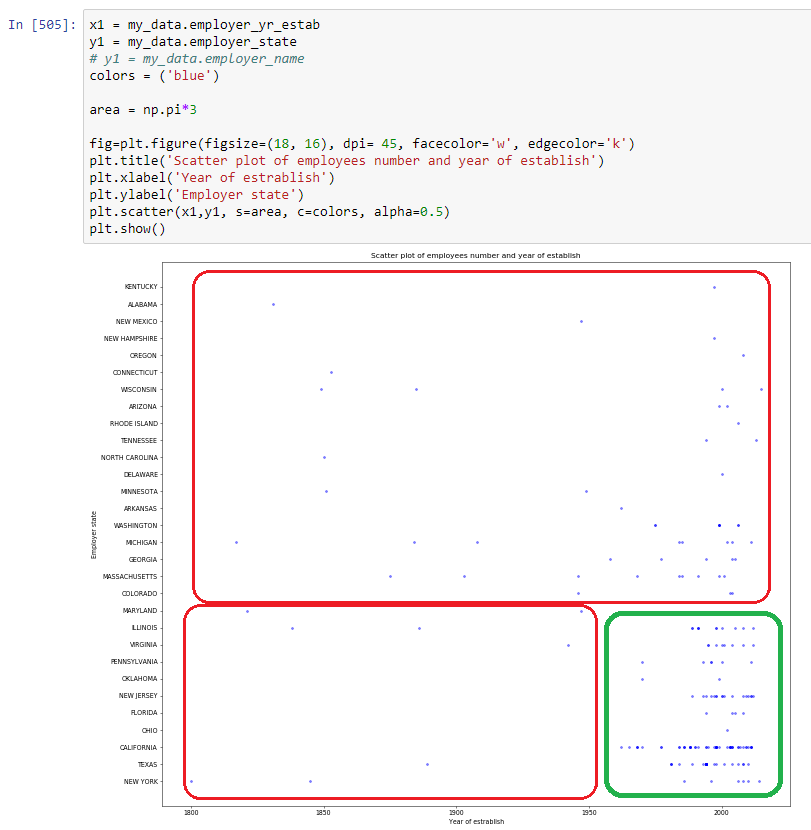


Figure 35: Scatter plot of employer state and employer year establish

### Employer number of employees and state

Another scatter plot is shown in figure 35 which is represented the selected randomly data sets between employer’s states and employees number. The green square identified the cluster data in this graph that has connection between employer state and employees number. In X-axis represent the employees number by individual amount and employer state in Y-axis. The main cluster is that amount of employee less than 200,000 and mostly is in California state. Another group is red squares which are identified the outliers, that represent the unusual data values that such as Virginia state has more than 1,400,000 employees.

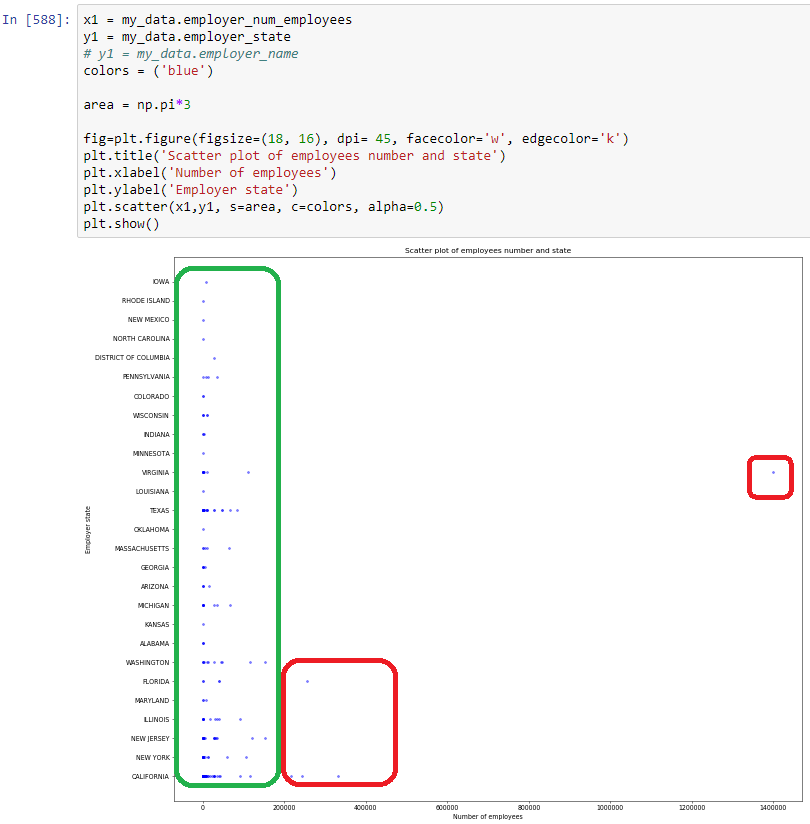


Figure 35: Scatter plot of employer state and employer num employees

### Interesting Finding in Attributes

During the exploring data set stage, some data has represented the interesting outcomes but one of the interesting attribute are ‘case received date’ and ‘decision date’ that assumed to be relate to each other by date. The frequently of values in each data set is vague and has a massive uniquely date, but if compare the values between them it has shown that mostly case values are mostly in year 2015 and 2016. The start date of case received rarely start since 2009 but decision date is start from 2014 to 2016. These date sets might represent the widely range of date, but it seems that most of the useful data are in lately year. In case received date mostly has data values as in 2015 years but the most decision date is within 2016 and also each year of case received date it mostly has decision date in next year, but it depends on case received date. For instant, if case received date is in the early 2015 sometimes the data values in decision date is also in year 2015 only just in some months later.

## Data pre-processing

The pre-processing is to prepare data set by using several techniques to re-format of the data in order to make data more readable and efficient use. Techniques such as binning, normalisation, discretisation, and binarization would be used. And the main tool will be used is Jupyter notebook and python libraries.

### Binning

Binning techniques has mainly two type which are ‘equi-width binning’ and ‘equi-depth binning’

#### Equi-width binning:

Firstly, applying equi-width binning technique to ‘employer\_num\_employees’ dataset will be performed in Jupyter notebook tool and required some following steps:

1. Import necessary libraries such as Pandas and CategoricalDtype from Pandas

2. Use Pandas to read csv file by using “kp2 = pd.read\_csv('Kunanon\_Pititheerachot.csv')” code.

3. Find range of Employee number by use “kp2.employer\_num\_employees.sort\_values(ascending=True)” code to rearrange dataset into ordering by ascending, and min number is 0 and max number is 2,200,000

4. Calculate bins from dividing 2,200,000 with 5 the result is 440,000 for each bin.

5. Use “employee\_num = ['0 - 440,000', '440001 - 880,000', '880,001 - 1,320,000', '1,320,001 -1,760,000', '1,760,001 - 2,200,000']” code to set range of each bin label

6. Use “kp2['employee\_number'] = pd.cut(kp2.employer\_num\_employees,

bins=[0, 440000, 880000, 1320000, 1760000, np.inf],

labels=employee\_num,

right=True).astype(CategoricalDtype(employee\_num, ordered=True))”

6.1 code to set up new column in kp2 (Kunanon\_Pititheerachot.csv), use Pandas cut function as pd.cut to identify each data value of employer\_num\_employees

6.2 identify by each bin using code “bins=[0, 440000, 880000, 1320000, 1760000, np.inf],”

6.3 create result column “labels=employee\_num”

7. Represent the result by using “kp2.sort\_values(by='employee\_number)'” code

8. And save result to csv file by “kp2.sort\_values(by='employee\_number').to\_csv("width\_binning.csv", index=False, encoding='utf8')”

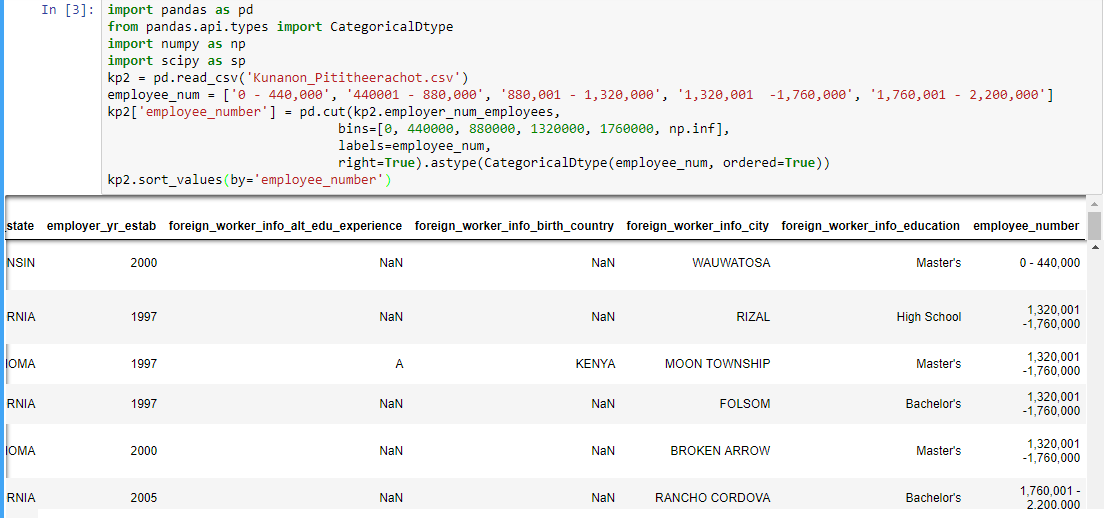


Figure 39: Equi-width binning technique.

#### Equi-depth binning:

In order to implement the Equi-depth binning technique is to follow the equi-width binning step but the different is to replace data value the using before with frequently of data instead. This technique requires some addition steps:

1. use

1.1 “data11 = kp.groupby('employer\_num\_employees').size()” – to count each occurrence in data values

1.2 “pd.options.display.max\_rows = None” – to shows every single row of occurrences

1.3 “data11.to\_csv("employer\_num\_employees\_size.csv", index=False, encoding='utf8')” – to create csv for storing number of occurrences

1.4 “data11” – to show result on screen

code to create new csv file to store number of employee and frequently of employee number data values (Figure 40).



Figure 40: Equi-depth binning technique.

2. use the same step as equi-width binning with min 1 and max 52 then use 5 bins such as 0, 10, 20, 30, 40.

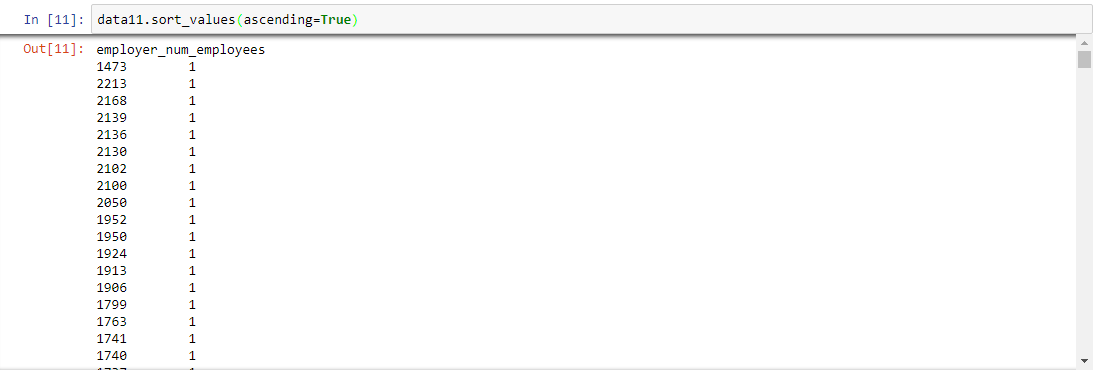


Figure 41: Equi-depth binning technique.

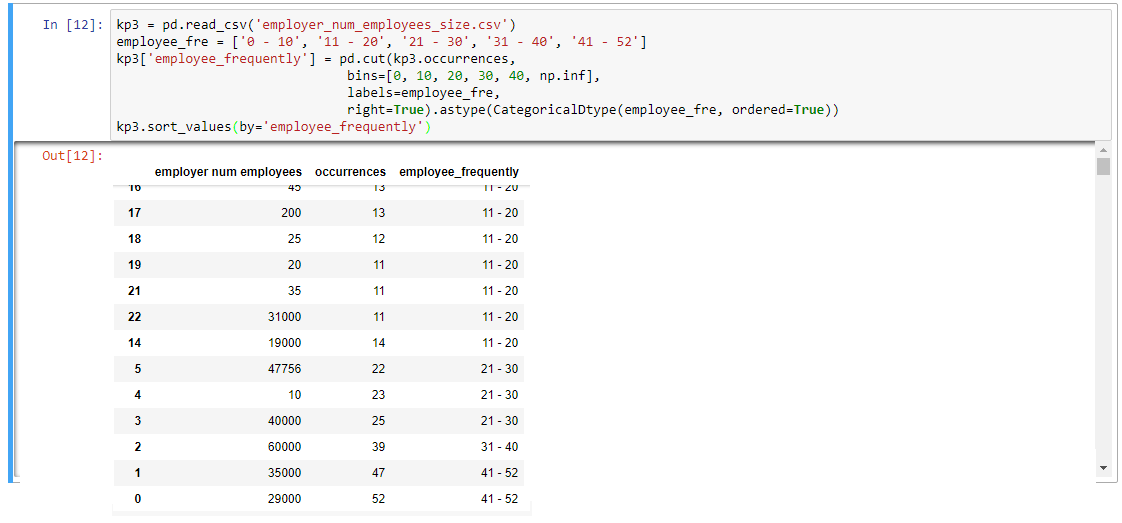


Figure 42: Equi-depth binning technique.

### Normalisation

#### Min/Max Normalisation:

Min/max normalisation is to convert values from dataset in to dataset range [0.0 – 1.0] in order to help visualisation more easier and much simpler to implement. This normalisation technique require some steps below:

1. To continue from binning technique I will new value to store csv file which is called “kp4” and next in order to find normalised value which is need to calculate with specific formula “A’ = (A -min) / (max - min) \* (new max – new min) – new min”, and already replaced with code as shown in figure 42 and 43.

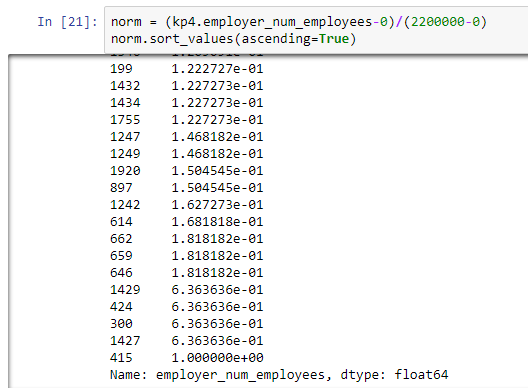


Figure 42: Min/max normalisation technique.



Figure 43: Min/max normalisation technique.

#### Z-Score Normalisation:

Z-score normalisation is used to help visualisation more easier and much simpler when the outliers, minimum, and maximum values are unknown by using mean and standard deviation instead. This normalisation technique requires some steps similar to min/max normalisation but different formula:

1. Formula is “A’ = (A - mean) / standard deviation” as shown below in figure 44, which is x\_bar is mean and sd is standard deviation and the result is shown in new column called “employee\_z\_score”.

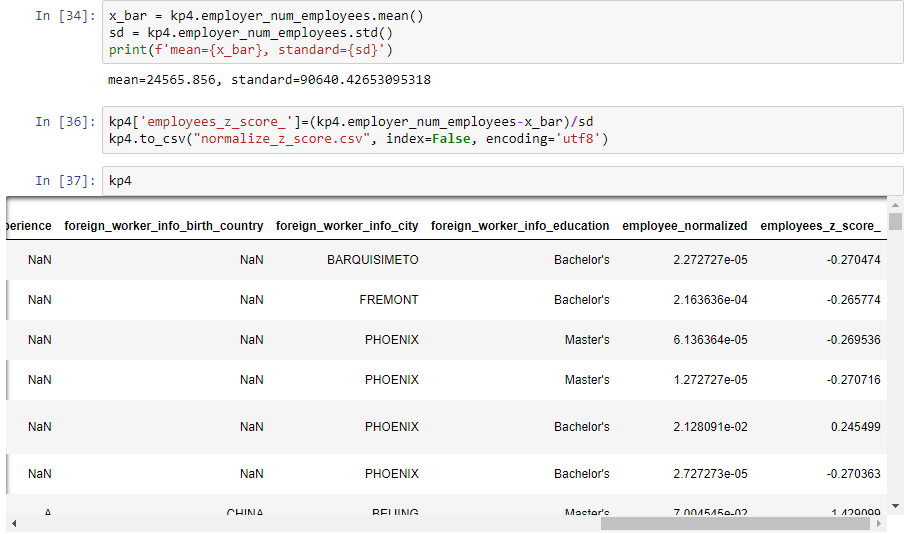


Figure 44: Z-score normalisation.

### Discretisation

To discretise the employer\_num\_employees attribute into the following categories:

Startup = 0-10;

Small\_Scale = 11-100;

Medium\_Scale = 101- 2000;

Large\_Scale = 2001-20000,

Giant\_Scale = 20001+.

This process require some step as binning technique to identify a number into 5 categories as mentioned with code shown in figure 45 in the first part then the code “groupby()” is applied to the current dataset groupby is used for scope dataset which is scoping employer\_scale dataset that created on the first part then use function “size()” to find occurrences or frequency that employer\_scale has occurred as a result in figure 45 the second part.

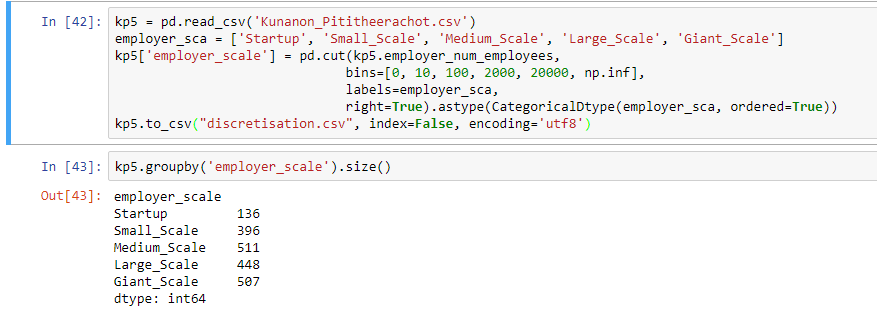


Figure 45: Discretisation technique.

### Binarisation

Binarisation technique is used for applying conversion of dataset’s values into binary number which are 0 and 1 identify by condition such as if in dataset foreign\_worker\_info\_education value is “Master’s” then return 1 else return 0. This technique benefit to constructing models such as decision tree model. Binarisation technique require small steps below:

1. Use pandas library to read csv file same as before then use lambda function from Pandas libraries as shown in upper part of figure 46 and create the condition such as “s” which is data value in education dataset if “s” equal to “Master’s” return 1 else 0.

2. Then apply new binaries values to the new column called “binarisation” as shown in bottom part in figure 46.

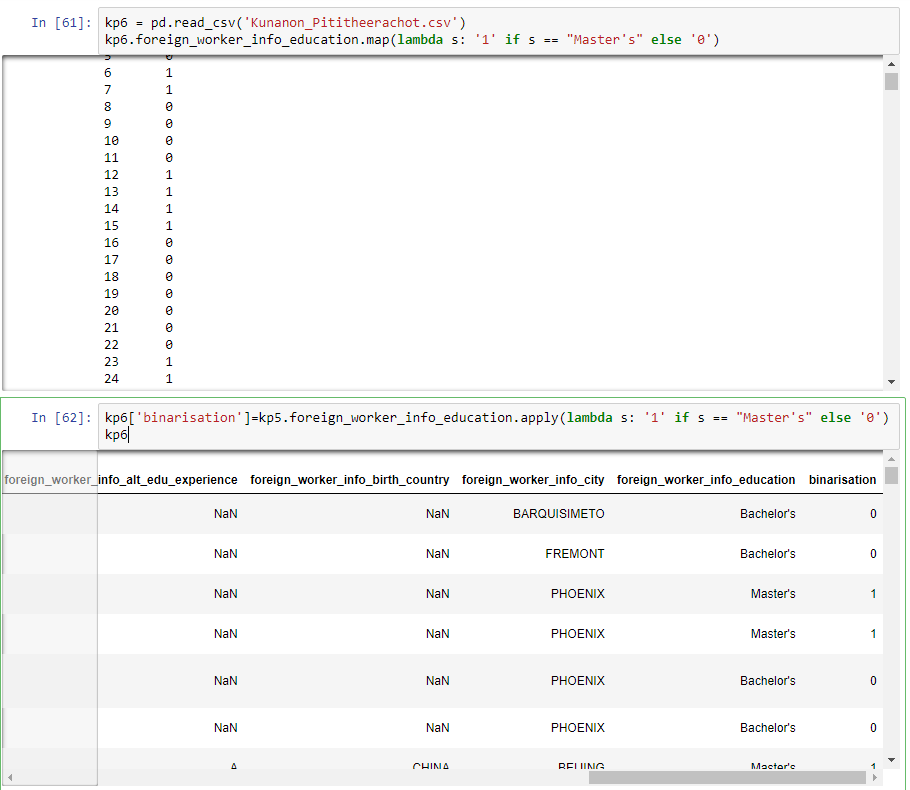


Figure 46: Binarisation technique.

## Summary

### Case received date

This attribute range start from 17/03/2009 to 16/12/2016. The most date are in year 2015 with 36% of total number of occurrences. The major cluster of values are start between early 2014 to 2016, and outliers are the rest (2009 to 3013) that combine all together less than 10% of total occurrences.

### Decision date

This attribute range start from date 01/10/2014 to 30/12/2016. The majority of values are clearly divide to two parts which are in year 2015 and 2016. And the in year 2016 has 1019 occurrences that 10% larger than 2015 that has 827 occurrences.

### Employer state

Employer state attribute has occurrences values between 1 to 536. The top three is California, Texas, and New Jersey. California has significant number of occurrences values than the second and the third place. The second place has only 290 occurrences which is a little bit more than half number of the first place, and the third place only has 160. This attribute could represent that most employer company is in California.

### Employer year establish

Employer year establish attribute is represent the majority of year which is range 1740 to 2016. As it shows that over the past 200 years employer mostly establish their business in year 1994 with 161 occurrences, then 1999 that has 110 occurrences and 1998 106 occurrences. According to attribute employer still establish their new business in every year even though the number is less than the past.

### Employer number employee

In this attribute represents the number of total employees of a business that has a huge range 0 to 2,200,000 employee. Although, the max is very high number, but average of employee is only 24,565 which mean very rare employer that has million employees. Mostly employers have about 5 digits number of employees, the most number of employees is 29,000 with 52 employers, 47 employers have 35,000, and 39 employers has 60,000 employees.

### Clusters and Outliers

### Employer Number Employees and Employer Year Establish

According to scatter plot graph above that mention about number of employees and establish year of employers that has significant rising of trend around 2000s, even though the number of employees does not that much but still provide an idea that in 21th centuries people have ideas about create new business more than before. In addition, the more employers the more employees followed and might have new trend for occupation and etc.

### Employer Year Establish and Employer State

As mention before about establish year that has a lot of new business around 2000s and the number of employees is has some relative. With this information could be related to where employers trend to do their business. In the cluster of employer establish year and employer state it has major interesting area which is California that has the around 500 of 2000 and also the year of establish still cluster same as number of employees which is mostly new business has occurred in 2000s with a little bit early or before.

### Conclusion

In conclusion, the analysation by exploring within all of the attributes are led to the idea that every processing stage of data analyse is necessary such as for cleaning data steps are important for exploration because the data come with some mess such as not in the properly same type, sometimes lack of values in some dataset, and etc. For another stage is data pre-processing that provide some interesting idea about information that can be transform into good shape in order to be used in any other use of data such as binarization can be done in order to make more accurate when use the dataset to create decision tree model. As a result, the most time is consumed by cleaning and exploring data to find and learn the data and after that can be consider is data useful or not and which one of them are truly useful.