英業達
Inventec

英業達 Inventec

Data Scientist

作者：劉建笙

目錄

英業達 Inventec

專案背景

英業達
Inventec

專案說明

共有15個自由欄位可以發揮

1. 分成三部分建立模型
   a. Clustering Model - *K-Means*
   b. Classification Model - *Decision Tree*
   c. Association Model - *Apriori Algorithm*
2. 以美國國會議案為主軸，運用三個模型進行分析
   a. K-Means: 探討五個國際關係主題
   b. Decision Tree: 分類議員們對支持FED升息的態度
   c. Association Model: 研究內政相關法案間的關聯性

英業達 Inventec

資料EDA

英業達
Inventec

EDA

資料型態

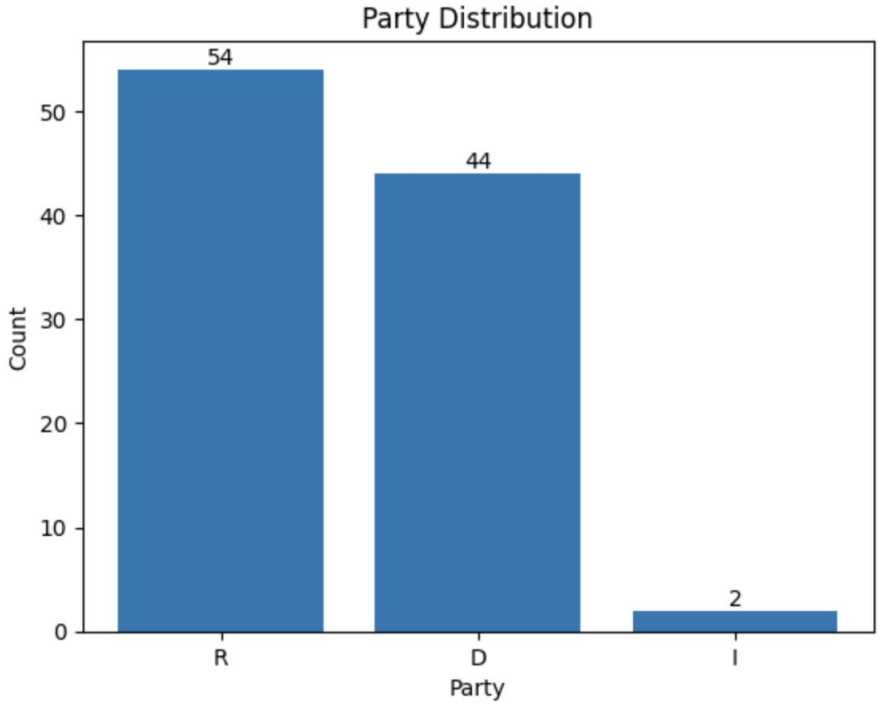
1. 3個非數值欄位
   a. 包含Name, Party, State
   b. 包含100個不同名字、3個不同政黨、50個不同州
2. 15個數值欄位
   a. 欄位名稱為數字1到15
   b. 需要自定義欄位1到15的意思
   c. 其中數值只包含0, 0.5, 1

EDA

政黨資料 EDA

共有三個政黨

1. R黨54人
2. D黨44人
3. I黨2人



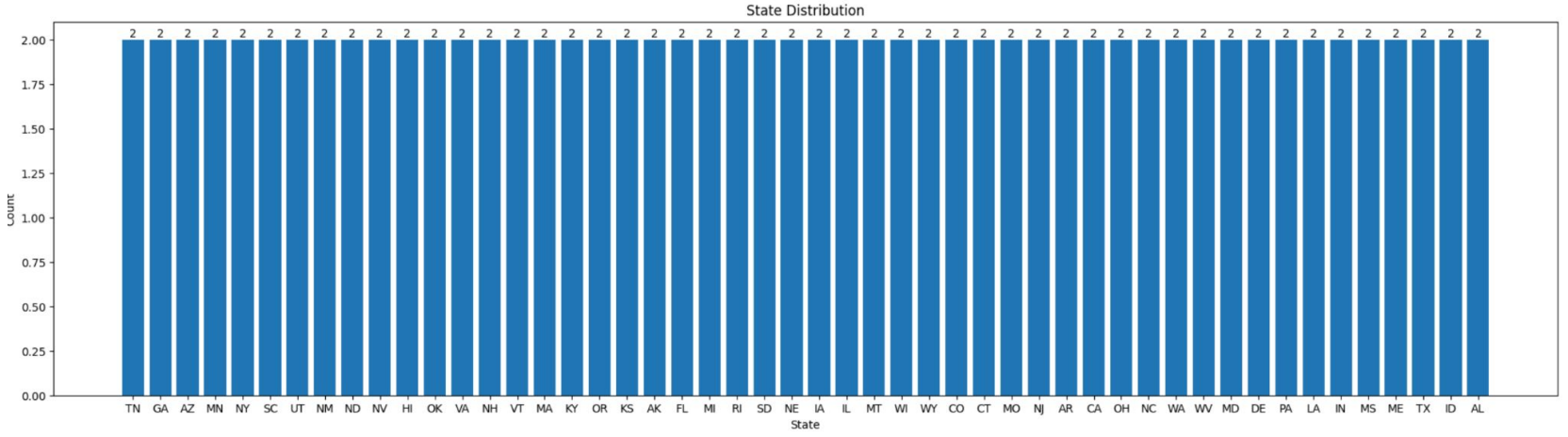
英業達
Inventec

EDA

英業達
Inventec

州資料 EDA

共有50個州，每個州各有兩位代表



EDA

整體資料 EDA

- 無空缺值
- Name, party, state資料為object
- 剩下資料為float

```
Data columns (total 18 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   name                        100 non-null    object
 1   party                       100 non-null    object
 2   state                       100 non-null    object
 3   armforce_in_middleeast      100 non-null    float64
 4   refugees_accpetance         100 non-null    float64
 5   china_5G_infra              100 non-null    float64
 6   china_trade_war             100 non-null    float64
 7   antitrust_laws              100 non-null    float64
 8   degree_level                100 non-null    float64
 9   assets_values               100 non-null    float64
 10  having_loans                100 non-null    float64
 11  immigrant                   100 non-null    float64
 12  support_or_not              100 non-null    float64
 13  support_abortion            100 non-null    float64
 14  support_free_childcare      100 non-null    float64
 15  support_free_education      100 non-null    float64
 16  support_1yr_maternity_leave 100 non-null    float64
 17  support_3yrs_parental_leave 100 non-null    float64
dtypes: float64(15), object(3)
memory usage: 14.2+ KB
```

英業達
Inventec

EDA

整體資料 EDA

- Categorical Data無重複值

```
Unique name: ['Alexander' 'Ayotte' 'Baldwin' 'Barrasso' 'Bennet' 'Blumenthal' 'Blunt'
 'Booker' 'Boozman' 'Boxer' 'Brown' 'Burr' 'Cantwell' 'Capito' 'Cardin'
 'Carper' 'Casey' 'Cassidy' 'Coats' 'Cochran' 'Collins' 'Coons' 'Corker'
 'Cornyn' 'Cotton' 'Crapo' 'Cruz' 'Daines' 'Donnelly' 'Durbin' 'Enzi'
 'Ernst' 'Feinstein' 'Fischer' 'Flake' 'Franken' 'Gardner' 'Gillibrand'
 'Graham' 'Grassley' 'Hatch' 'Heinrich' 'Heitkamp' 'Heller' 'Hirono'
 'Hoeven' 'Inhofe' 'Isakson' 'Johnson' 'Kaine' 'King' 'Kirk' 'Klobuchar'
 'Lankford' 'Leahy' 'Lee' 'Manchin' 'Markey' 'McCain' 'McCaskill'
 'McConnell' 'Menendez' 'Merkley' 'Mikulski' 'Moran' 'Murkowski' 'Murphy'
 'Murray' 'Nelson' 'Paul' 'Perdue' 'Peters' 'Portman' 'Reed' 'Reid'
 'Risch' 'Roberts' 'Rounds' 'Rubio' 'Sanders' 'Sasse' 'Schatz' 'Schumer'
 'Scott' 'Sessions' 'Shaheen' 'Shelby' 'Stabenow' 'Sullivan' 'Tester'
 'Thune' 'Tillis' 'Toomey' 'Udall' 'Vitter' 'Warner' 'Warren' 'Whitehouse'
 'Wicker' 'Wyden']
Number of unique name: 100

Unique state: ['TN' 'NH' 'WI' 'WY' 'CO' 'CT' 'MO' 'NJ' 'AR' 'CA' 'OH' 'NC' 'WA' 'WV'
 'MD' 'DE' 'PA' 'LA' 'IN' 'MS' 'ME' 'TX' 'ID' 'MT' 'IL' 'IA' 'NE' 'AZ'
 'MN' 'NY' 'SC' 'UT' 'NM' 'ND' 'NV' 'HI' 'OK' 'GA' 'VA' 'VT' 'MA' 'KY'
 'OR' 'KS' 'AK' 'FL' 'MI' 'RI' 'SD' 'AL']
Number of unique state: 50

Unique party: ['R' 'D' 'I']
Number of unique party: 3
```

英業達 Inventec

Clustering

K-Means

英業達
Inventec

國際關係資料定義

英業達 Inventec

- armforce_in_middleeast: 美國是否應該支持中東駐紮軍隊？
- refugees_acceptance: 美國是否應該接受難民？
- china_5G_infra: 美國是否應該使用中國進口的5G基地台？
- trade_war: 對於中美貿易戰的態度？
- antitrust_law: 對於反托拉斯法的態度？

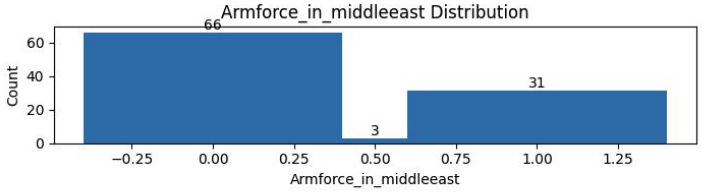
欄位值解說

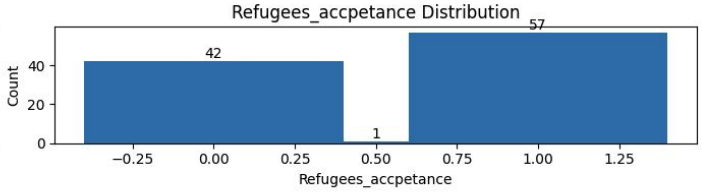
- 1.0 : 贊成
- 0.5 : 中立
- 0.0 : 反對

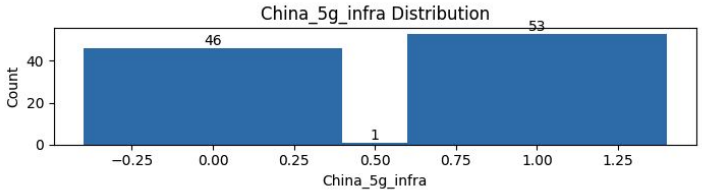
	name	party	state	armforce_in_middleeast	refugees_accpetance	china_5G_infra	china_trade_war	antitrust_laws
0	Alexander	R	TN	0.0	1.0	1.0	1.0	1.0
1	Ayotte	R	NH	0.0	1.0	1.0	1.0	1.0
2	Baldwin	D	WI	1.0	0.0	0.0	1.0	0.0
3	Barrasso	R	WY	0.0	1.0	1.0	1.0	1.0
4	Bennet	D	CO	0.0	0.0	0.0	1.0	0.0

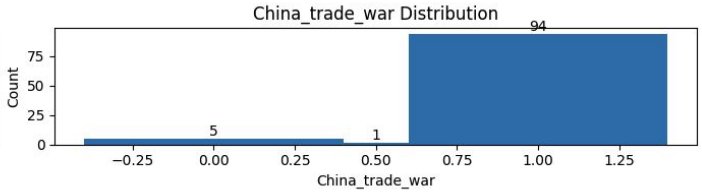
EDA

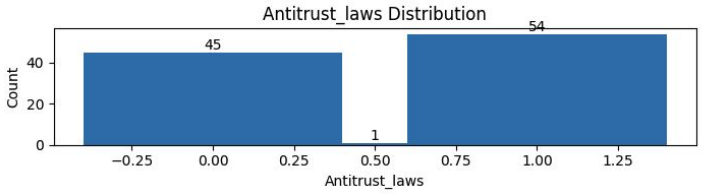
國際關係資料 EDA







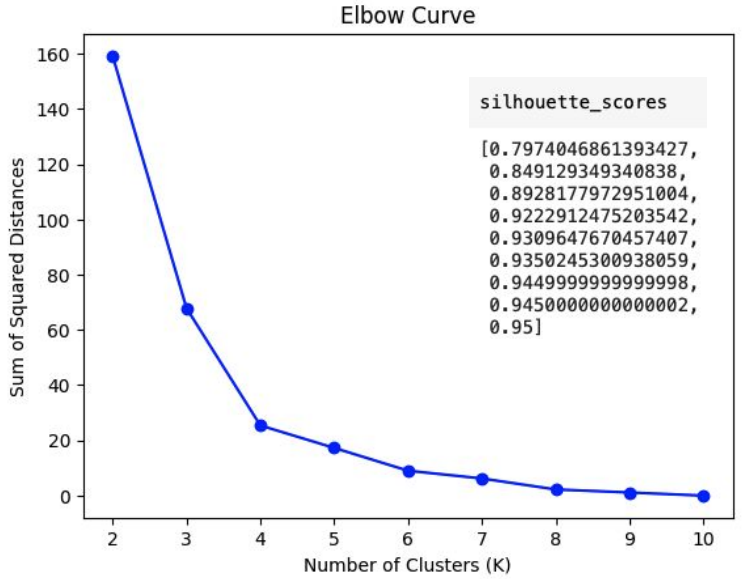




英業達
Inventec

K-Means

1. 創建只含有Numerical Values的DataFrame
   a. 只存儲armforce_in_middleeast, refugees_acceptance, china_5G_infra, trade_war, antitrust_law
2. 檢查資料是否需要進行縮放
   a. 運用StandardScaler，最終不需要
      i. 特徵資料間的距離皆相等
      ii. 特徵間的資料單位皆相同
3. 運用Elbow Curve以及的Silhouette尋找最佳的K
   a. 最佳的Silhouette分數落在10，但10有點不合理
   b. 運用Elbow Curve找到K=4
4. 進行PCA縮放
   a. 降至2維空間，減少特徵數量
5. 分別建立K=4以及K=10的模型
6. 寫進DataFrame內，查看每個人是在哪個Cluster



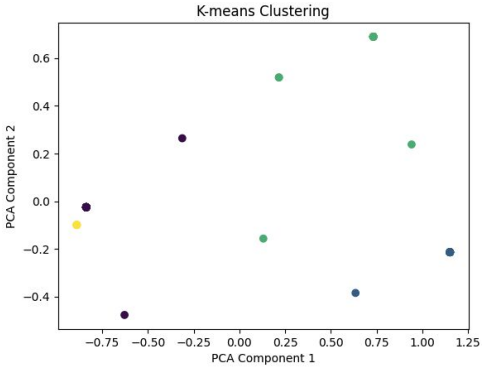
K-Means

K=10

- 大部分人多集中於前面5個Cluster

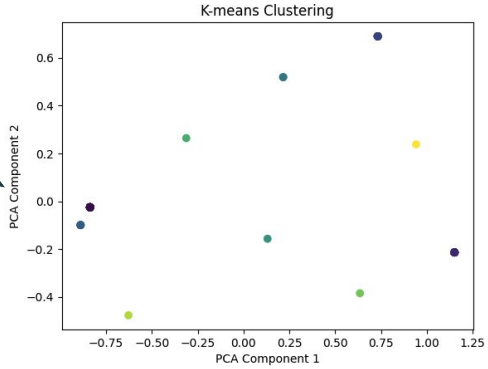
K=4

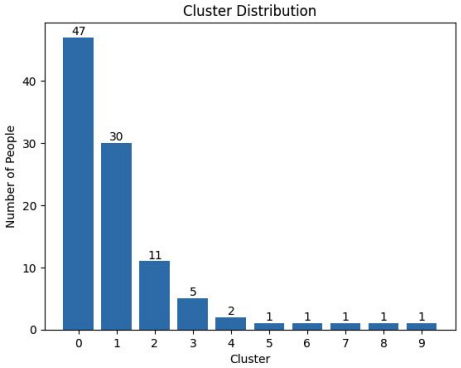
- 資料相較均勻地被歸納

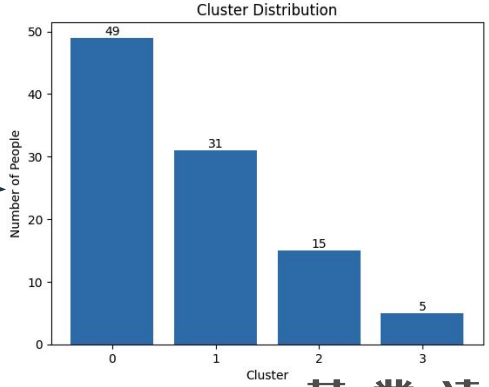


結論

- K=4較適合
  - K=10到第五群開始數量極度少，之間相差不大







英業達
Inventec

K-Means

最終產出

- 每個人屬於哪一群皆寫進DataFrame

	name	party	state	armforce_in_middleeast	refugees_accpetance	china_5G_infra	china_trade_war	antitrust_laws	Cluster
0	Alexander	R	TN	0.0	1.0	1.0	1.0	1.0	0
1	Ayotte	R	NH	0.0	1.0	1.0	1.0	1.0	0
2	Baldwin	D	WI	1.0	0.0	0.0	1.0	0.0	1
3	Barrasso	R	WY	0.0	1.0	1.0	1.0	1.0	0
4	Bennet	D	CO	0.0	0.0	0.0	1.0	0.0	2

Classification

Decision Tree

英業達
Inventec

是否支持升息的資料定義

- degree_level: 教育程度
- assets_values: 財產價值
- having_loans: 擁有貸款的金額
- immigrant: 移民背景調查
- support_or_not: 是否贊同FED升息？

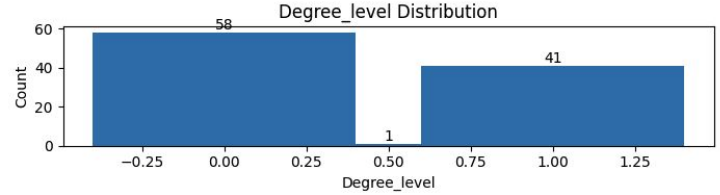
- 0.0 - 學士學位、0.5 - 碩士學位、1.0 - 博士學位
- 0.0 - $10W以下、0.5 - $10W-$50W、1.0 - $50W以上
- 0.0 - 無貸款、0.5 - $1-$20W、1.0 - $20W以上
- 0.0 - 無移民背景、0.5 - 第一代新移民、1.0 - 第二或以上代移民
- (0.0)No - 反對、(0.5)Neutral - 不表態/沒意見/中性、(1.0)Yes - 贊成

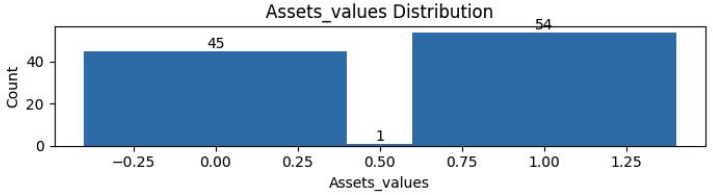
	name	party	state	degree_level	assets_values	having_loans	immigrant	support_or_not
0	Alexander	R	TN	0.0	0.0	1.0	1.0	Yes
1	Ayotte	R	NH	0.0	0.0	1.0	0.0	Yes
2	Baldwin	D	WI	1.0	0.0	1.0	0.0	No
3	Barrasso	R	WY	0.0	1.0	1.0	1.0	Yes
4	Bennet	D	CO	1.0	0.0	1.0	0.0	No

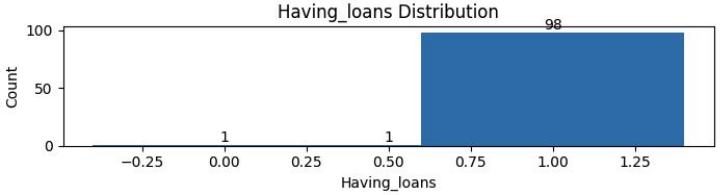
英業達 Inventec

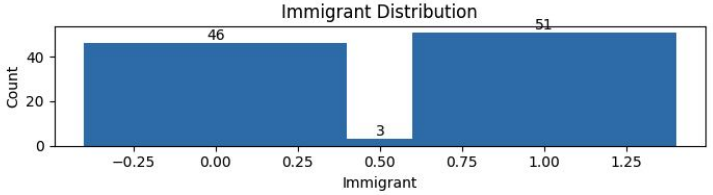
EDA

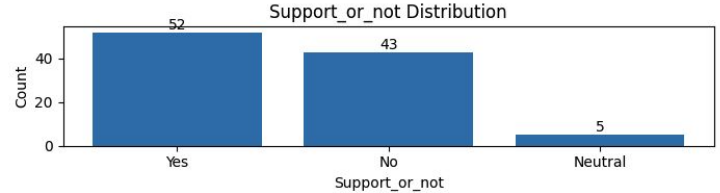
是否支持升息的資料 EDA











英業達
Inventec

Decision Tree

1. 將support_or_not變數的值都先轉為No, Neutral, Yes
2. 創建不含有姓名的DataFrame
   a. 因為姓名在決策樹中不重要
3. 將資料DataFrame分成X以及y並將訓資料轉為Dummies
   a. X用來當作模型訓練以及預測的Input
   b. y用來訓練以及預測的目標變數
   c. 針對政黨、地區等兩個Categorical資料轉換為Dummies
4. 分為訓練集80%以及測試集20%
5. 模型擬合、訓練並進行測試預測
6. 利用測試集獲取Accuracy、Precision、Recall以及F1分數
7. 把樹給視覺化出來
   a. 但是MacBook沒辦法用

```
Accuracy: 0.95
Precision: 1.0
Recall: 0.95
F1 Score: 0.9741379310344829
```

英業達 Inventec

Decision Tree

（接上頁）

1. 利用測試集獲取Accuracy、Precision、Recall以及F1分數
2. 自行創建一筆資料真實進行預測
   a. 根據資料獲得中性/不表態的結果

```
Accuracy: 0.95
Precision: 1.0
Recall: 0.95
F1 Score: 0.9741379310344829
```

```
Name: Morris
Party: R
State: SC
Degreee Level: 1.0
Assest Amount: 1.0
Haveing Loans: 0.0
Immigrant: 0.5
```

```
unseen_data = pd.DataFrame({'party': ['R'],
                            'state': ['SC'],
                            'degree_level': [1.0],
                            'assets_amount': [1.0],
                            'having_loans': [0.0],
                            'immigrant': [0.5]})

unseen_data_encoded = pd.get_dummies(unseen_data)

unseen_data_encoded = unseen_data_encoded.reindex(columns=X.columns, fill_value=0)

prediction = model.predict(unseen_data_encoded)

print("Morris's stance toward the increase of interest rate is", prediction[0])
```

```
Morris's stance toward the increase of interest rate is Neutral
```

英業達 Inventec

Association Analysis

Apriori Algorithm

英業達
Inventec

內政法案關聯性研究資料定義

英業達 Inventec

- support_abortion: 支持墮胎的立場
- support_free_childcare: 支持免費嬰兒照護的立場
- support_free_education: 支持免費教育
- support_2yrs_maternity_leave: 支持兩年產假
- support_6yrs_parental_leave: 支持六年孩童陪伴假

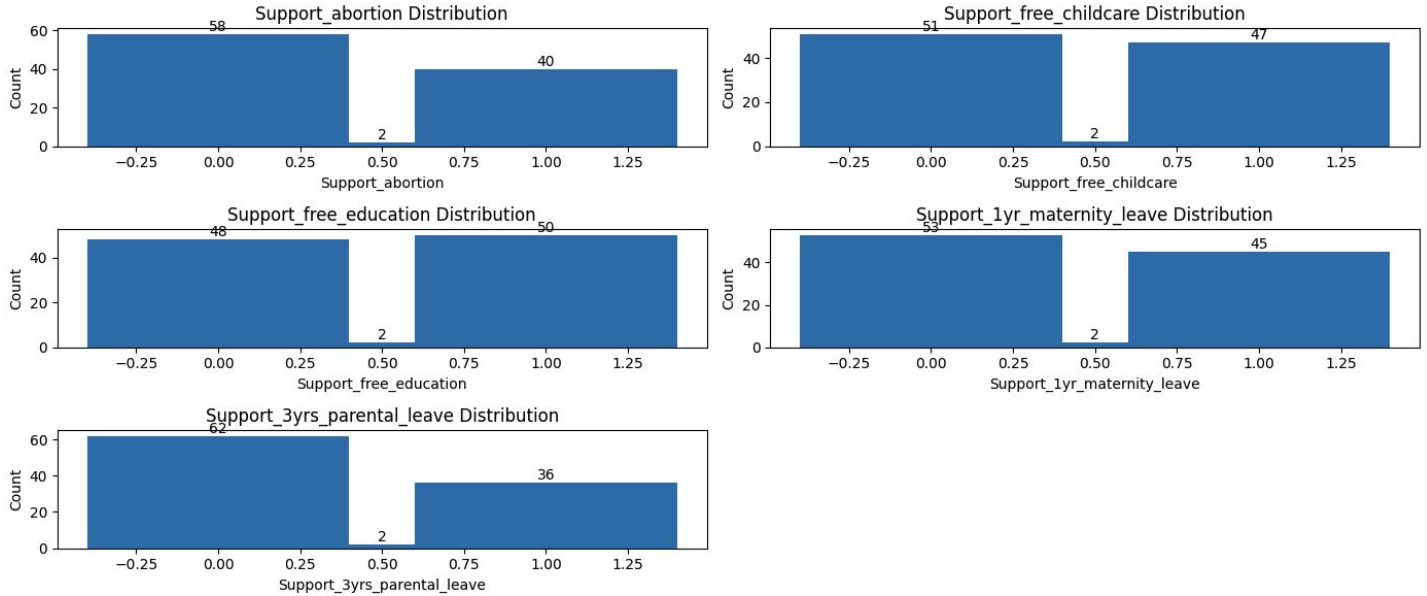
欄位值解說

- 1.0 : 贊成
- 0.5 : 中立
- 0.0 : 反對

	name	party	state	support_abortion	support_free_childcare	support_free_education	support_1yr_maternity_leave	support_3yrs_parental_leave
0	Alexander	R	TN	0.0	0.0	0.0	0.0	0.0
1	Ayotte	R	NH	0.0	1.0	0.0	1.0	0.0
2	Baldwin	D	WI	1.0	1.0	0.0	1.0	1.0
3	Barrasso	R	WY	0.0	0.0	1.0	0.0	0.0
4	Bennet	D	CO	0.0	1.0	0.0	1.0	0.0

EDA

內政法案關聯性研究資料 EDA



英業達
Inventec

Apriori Algorithm

英業達 Inventec

1. 創建不含有姓名、政黨、州等 Categorical Data的DataFrame
2. 將意見為中性（0.5）的Data Row刪除
   a. 對法案表決，有時候支持有時候不支持或是部分支持部分不支持是不合理的
   b. 中性意見等於沒意見，無助於模型學習
3. 模型運用Apriori演算法擬合
   a. 最小的Support設定為0.2
4. 產生關聯規則
5. 獲取Support, Confidence, Lift分數並且排序

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
46	(support_free_childcare)	(support_3yrs_parental_leave, support_1yr_mate...	0.479592	0.346939	0.346939	0.723404	2.085106	0.180550	2.361068
33	(support_free_childcare)	(support_3yrs_parental_leave, support_1yr_mate...	0.479592	0.357143	0.357143	0.744681	2.085106	0.185860	2.517857
21	(support_free_childcare)	(support_3yrs_parental_leave, support_abortion)	0.479592	0.357143	0.357143	0.744681	2.085106	0.185860	2.517857
47	(support_1yr_maternity_leave)	(support_free_childcare, support_3yrs_parental...	0.459184	0.357143	0.346939	0.755556	2.115556	0.182945	2.629870
27	(support_1yr_maternity_leave)	(support_3yrs_parental_leave, support_abortion)	0.459184	0.357143	0.346939	0.755556	2.115556	0.182945	2.629870