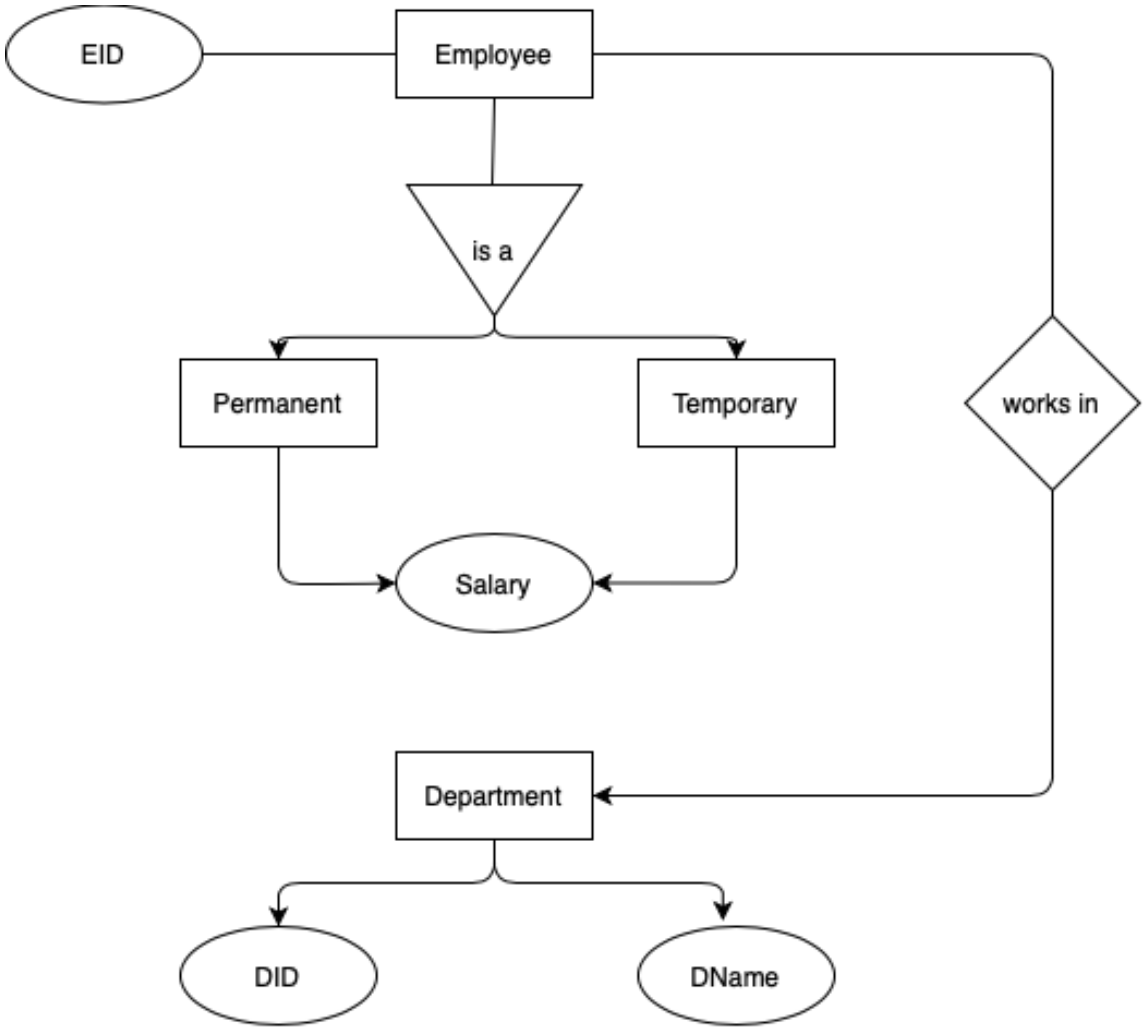
Course Code: CSL603	**Course Name: DWM LAB**
Class: TE-CO	**Batch: 3**
Roll no: 18CO63	**Name: SHAIKH TAUSEEF MUSHTAQUE ALI**

Experiment :01

Aim: Create an EER Model for a given case study.

Output:

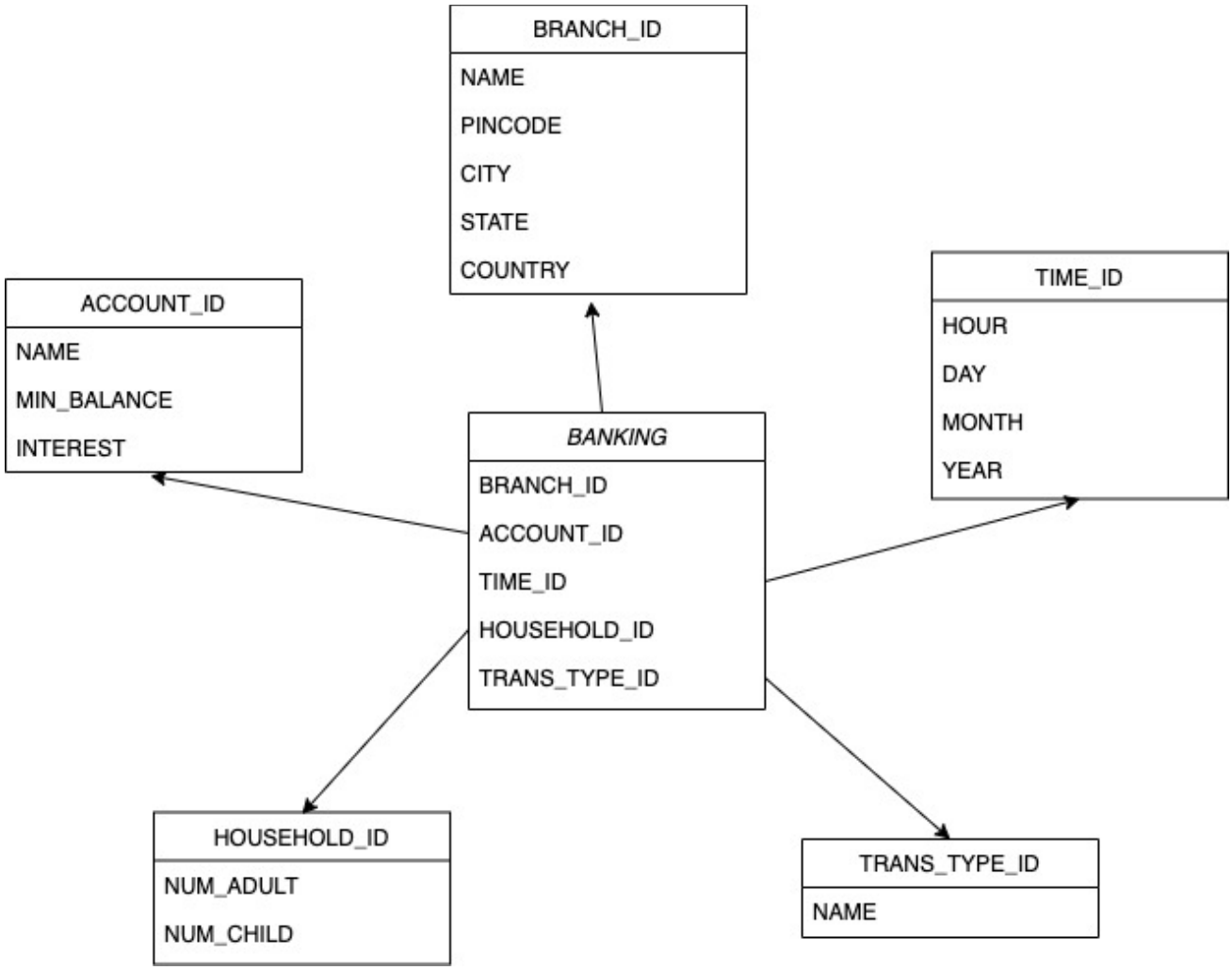


Course Code: CSL603	Course Name: DWM LAB
Class: TE-CO	Batch: 3
Roll no: 18CO63	Name: SHAIKH TAUSEEF MUSHTAQUE ALI

Experiment :02

Aim: Draw StarSchema for the said topic.

Output:



Course Code: CSL603	**Course Name: DWM LAB**
Class: TE-CO	**Batch: 3**
Roll no: 18CO63	**Name: SHAIKH TAUSEEF MUSHTAQUE ALI**

Experiment :03

Aim: Draw OLAP operations for the given case study.

Data:

P.T.O

Q] Consider a Data Warehouse for a hospital where there are three dimension
(a) Doctor (b) Patient (c) Time
And two measures i) count ii) charge where charge is the fee that the doctor charges a patient for a visit
Using the above example describe the following OLAP operations.
1) Slice 2) Dice 3) Roll up 4) Drill down
5) Pivot

⇒ Dimention Table:

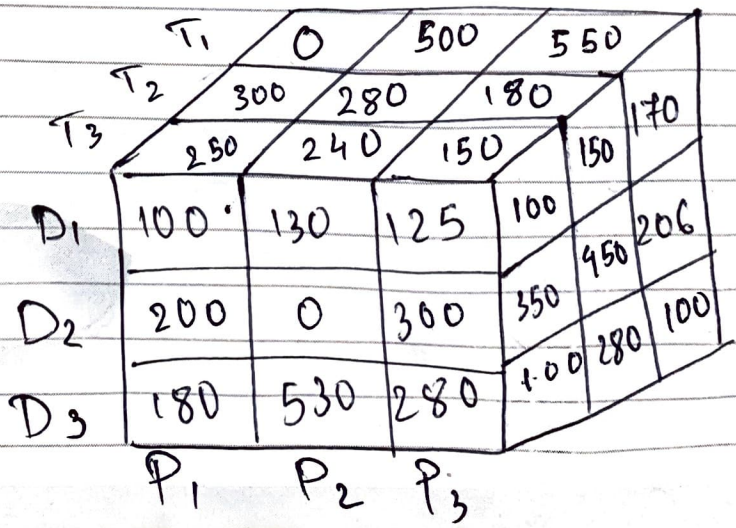
1) Doctor (DID, name, mob, add, specialization)

2) Patient (PID, name, mob, add)

3) Time (TID, day, month, quarter, year)

Fact Table:

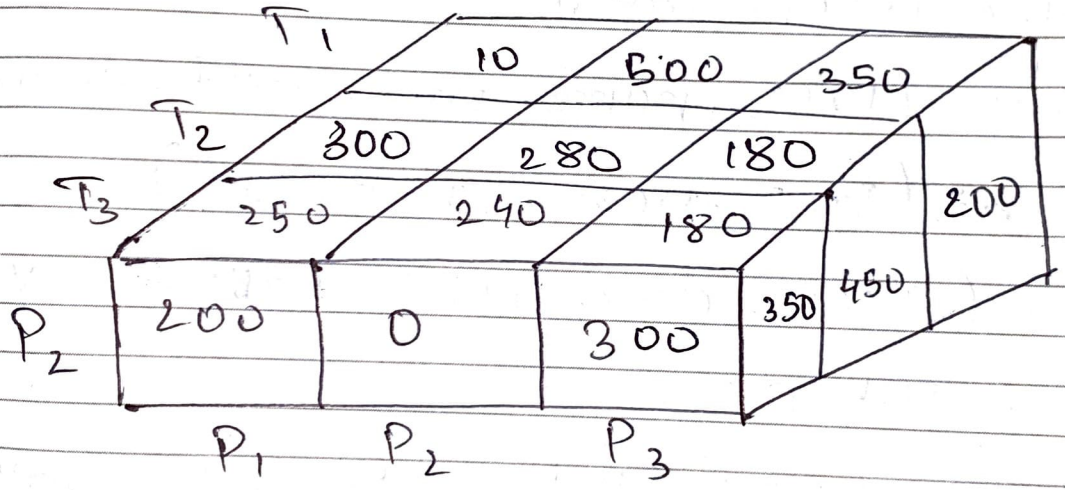
fact_table (DID, PID, TID, count, charge)



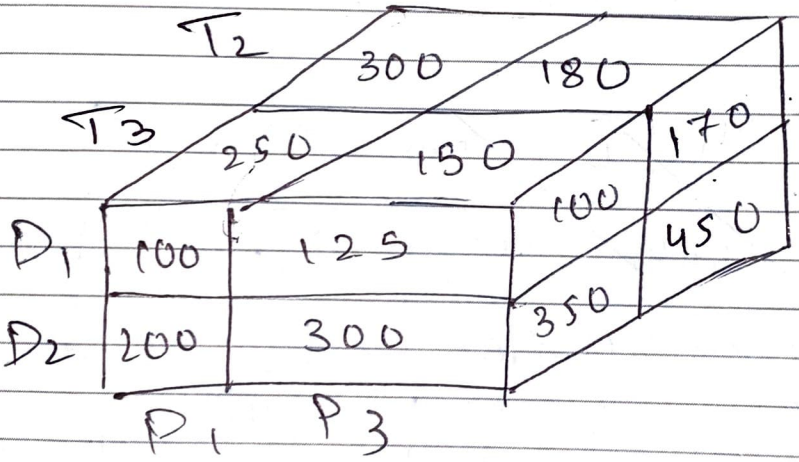
Operation 3:

1) Slice:

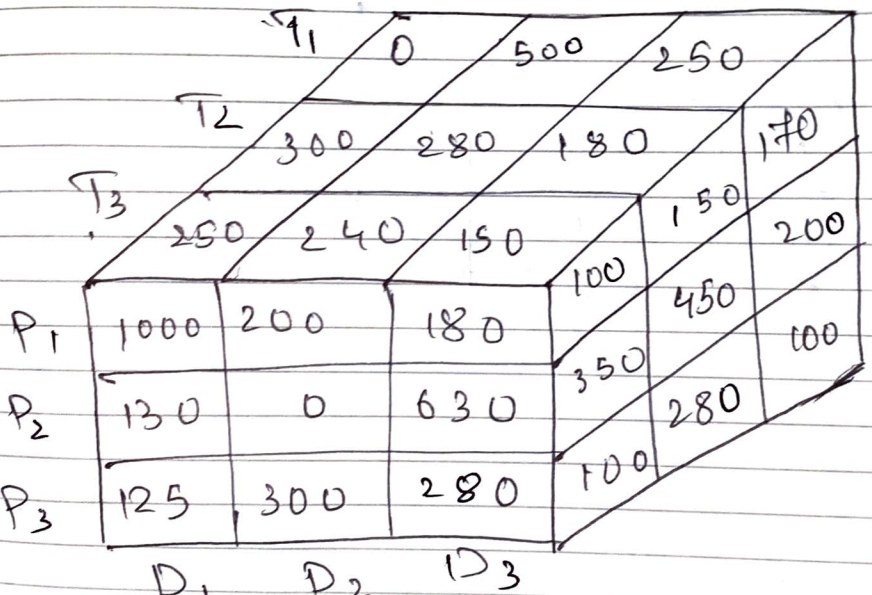
Slice on fact table with DID = 1 this acts the cube at DID = 2 along the time & Patient axis it will display area of cube in which time on x & patient on y axis



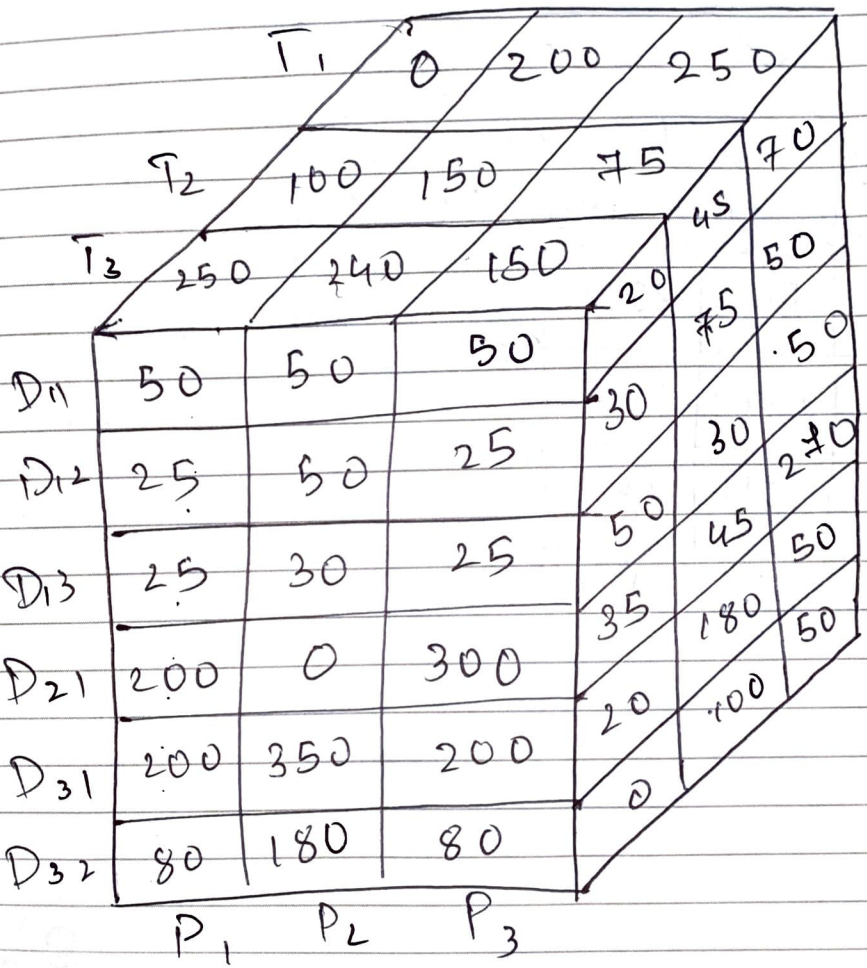
2) Dice: It is a subcube of main cube Thus it cuts the cube with more than predicate like dice on cube with DID=2 & DID=1 & PID=1 & PID=3 & TID= 02 & 03



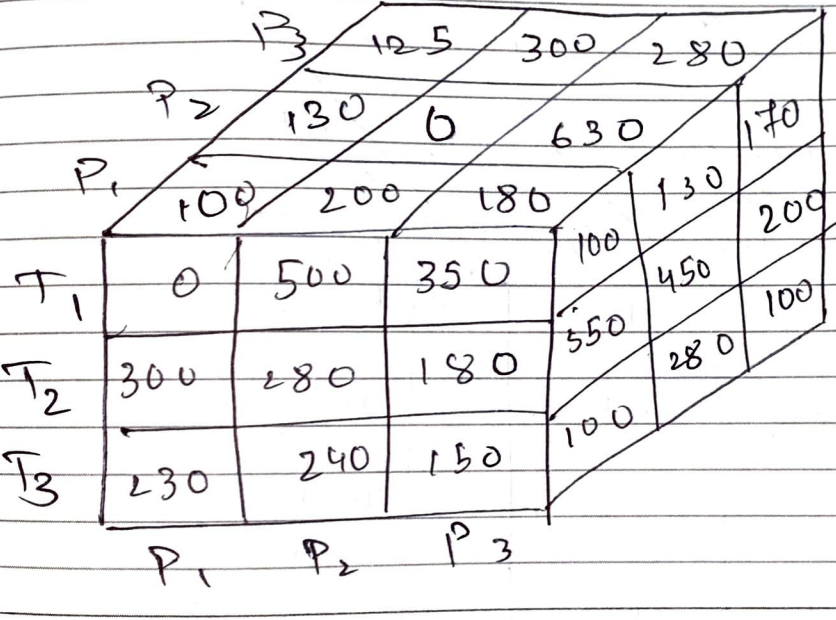
3) Roll up: It gives summary based on concept hierarchi or. Assuming there exists concept hierarchy in patient table as state → city → Location. The roll up will summarize the changes or count in terms of city or further roll up will give changes for a particular state etc.



4) Drill Down: It is opposite to roll up that means if currently cube is summarised with also show detailed view.



5) Pivot: It rotates the cube sub cube or rolled up or drilled down cube, thus changing the view of the cube.



Course Code: CSL603	**Course Name: DWM LAB**
Class: TE-CO	**Batch: 3**
Roll no: 18CO63	**Name: SHAIKH TAUSEEF MUSHTAQUE ALI**

Experiment :04

Aim: To perform various OLAP operations such as slice, dice, drilldown, rollup, pivot

Output:

Setting environment for using XAMPP for Windows.

tp@DESKTOP-K5A8CTB e:\xampp

mysql -u root -p

Enter password:

Welcome to the MariaDB monitor. Commands end with ; or \g.

Your MariaDB connection id is 108

Server version: 10.4.18-MariaDB mariadb.org binary distribution

Copyright (c) 2000, 2018, Oracle, MariaDB Corporation Ab and others.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

MariaDB [(none)]> use dwm;

Database changed

MariaDB [dwm]> show tables;

```
+---------------+
| Tables_in_dwm |
+---------------+
| abc           |
| fact_sales    |
+---------------+
```

2 rows in set (0.001 sec)

MariaDB [dwm]> select * from fact_sales;

id	year	month	product	sales_qty
1	2015	1	hard disk	15000
2	2015	2	hard disk	9000
3	2015	3	hard disk	16000
4	2015	4	hard disk	1000
5	2015	5	hard disk	9690
6	2015	6	hard disk	7800
7	2015	7	hard disk	13000
8	2015	8	hard disk	9800
9	2015	9	hard disk	13000
10	2015	10	hard disk	12680
11	2015	11	hard disk	15690
12	2015	12	hard disk	14200
13	2016	1	hard disk	1200
14	2016	2	hard disk	12200
15	2016	3	hard disk	1900
16	2016	4	hard disk	2500
17	2016	5	hard disk	15000
18	2016	6	hard disk	11000
19	2016	7	hard disk	10890
20	2016	8	hard disk	7300
21	2016	9	hard disk	9560
22	2016	10	hard disk	10610
23	2016	11	hard disk	19830
24	2016	12	hard disk	5630
25	2017	1	hard disk	15302
26	2017	2	hard disk	13090
27	2017	3	hard disk	11890
28	2017	4	hard disk	18722

29	2017	5	hard disk	8780
30	2017	6	hard disk	9820
31	2017	7	hard disk	3400
32	2017	8	hard disk	3190
33	2017	9	hard disk	8800
34	2017	10	hard disk	12920
35	2017	11	hard disk	1920
36	2017	12	hard disk	13050
37	2018	1	hard disk	1090
38	2018	2	hard disk	2908
39	2018	3	hard disk	11800
40	2018	4	hard disk	13900
41	2018	5	hard disk	9900
42	2018	6	hard disk	8350
43	2018	7	hard disk	5650
44	2018	8	hard disk	4950
45	2018	9	hard disk	9670
46	2018	10	hard disk	8020
47	2018	11	hard disk	10290
48	2018	12	hard disk	12030
49	2015	1	keyboard	1230
50	2015	2	keyboard	3000
51	2015	3	keyboard	3400
52	2015	4	keyboard	6700
53	2015	5	keyboard	1790
54	2015	6	keyboard	12900
55	2015	7	keyboard	2980
56	2015	8	keyboard	10900
57	2015	9	keyboard	9980
58	2015	10	keyboard	7880
59	2015	11	keyboard	6890
60	2015	12	keyboard	11890

61	2016	1	keyboard	8090
62	2016	2	keyboard	9010
63	2016	3	keyboard	6570
64	2016	4	keyboard	6100
65	2016	5	keyboard	7768
66	2016	6	keyboard	1908
67	2016	7	keyboard	11900
68	2016	8	keyboard	8900
69	2016	9	keyboard	8766
70	2016	10	keyboard	9878
71	2016	11	keyboard	7800
72	2016	12	keyboard	9000
73	2017	1	keyboard	8000
74	2017	2	keyboard	10900
75	2017	3	keyboard	9870
76	2017	4	keyboard	1770
77	2017	5	keyboard	7170
78	2017	6	keyboard	9880
79	2017	7	keyboard	10880
80	2017	8	keyboard	10180
81	2017	9	keyboard	12000
82	2017	10	keyboard	10200
83	2017	11	keyboard	12350
84	2017	12	keyboard	11310
85	2018	1	keyboard	9289
86	2018	2	keyboard	9010
87	2018	3	keyboard	1920
88	2018	4	keyboard	11920
89	2018	5	keyboard	9080
90	2018	6	keyboard	9000
91	2018	7	keyboard	8270
92	2018	8	keyboard	4220

93	2018	9	keyboard	2230
94	2018	10	keyboard	6770
95	2018	11	keyboard	7879
96	2018	12	keyboard	8089
97	2015	1	monitor	7009
98	2015	2	monitor	8999
99	2015	3	monitor	8780
100	2015	4	monitor	1780
101	2015	5	monitor	1080
102	2015	6	monitor	2290
103	2015	7	monitor	8290
104	2015	8	monitor	7090
105	2015	9	monitor	8780
106	2015	10	monitor	6760
107	2015	11	monitor	3029
108	2015	12	monitor	6780
109	2016	1	monitor	2290
110	2016	2	monitor	8900
111	2016	3	monitor	7000
112	2016	4	monitor	2800
113	2016	5	monitor	8020
114	2016	6	monitor	7878
115	2016	7	monitor	9800
116	2016	8	monitor	8950
117	2016	9	monitor	1090
118	2016	10	monitor	3490
119	2016	11	monitor	13490
120	2016	12	monitor	12390
121	2017	1	monitor	9180
122	2017	2	monitor	8000
123	2017	3	monitor	7890
124	2017	4	monitor	3890

125	2017	5	monitor	2990
126	2017	6	monitor	6990
127	2017	7	monitor	12000
128	2017	8	monitor	11020
129	2017	9	monitor	9800
130	2017	10	monitor	7800
131	2017	11	monitor	9878
132	2017	12	monitor	9800
133	2018	1	monitor	8700
134	2018	2	monitor	8670
135	2018	3	monitor	6770
136	2018	4	monitor	8790
137	2018	5	monitor	1980
138	2018	6	monitor	7789
139	2018	7	monitor	9880
140	2018	8	monitor	10890
141	2018	9	monitor	7677
142	2018	10	monitor	8700
143	2018	11	monitor	6750
144	2018	12	monitor	8700
145	2015	1	pendrive	18700
146	2015	2	pendrive	12900
147	2015	3	pendrive	19920
148	2015	4	pendrive	10190
149	2015	5	pendrive	11900
150	2015	6	pendrive	11800
151	2015	7	pendrive	18300
152	2015	8	pendrive	15670
153	2015	9	pendrive	12800
154	2015	10	pendrive	11900
155	2015	11	pendrive	12000
156	2015	12	pendrive	9010

157	2016	1	pendrive	13440
158	2016	2	pendrive	7040
159	2016	3	pendrive	16010
160	2016	4	pendrive	14500
161	2016	5	pendrive	12500
162	2016	6	pendrive	12920
163	2016	7	pendrive	11095
164	2016	8	pendrive	13080
165	2016	9	pendrive	12900
166	2016	10	pendrive	11200
167	2016	11	pendrive	12800
168	2016	12	pendrive	17800
169	2017	1	pendrive	19800
170	2017	2	pendrive	11900
171	2017	3	pendrive	9500
172	2017	4	pendrive	15000
173	2017	5	pendrive	17020
174	2017	6	pendrive	10220
175	2017	7	pendrive	11230
176	2017	8	pendrive	18050
177	2017	9	pendrive	12350
178	2017	10	pendrive	12350
179	2017	11	pendrive	10220
180	2017	12	pendrive	17620
181	2018	1	pendrive	16000
182	2018	2	pendrive	18720
183	2018	3	pendrive	12902
184	2018	4	pendrive	15020
185	2018	5	pendrive	17000
186	2018	6	pendrive	13900
187	2018	7	pendrive	12330
188	2018	8	pendrive	17800

189	2018	9	pendrive	15400
190	2018	10	pendrive	11200
191	2018	11	pendrive	11320
192	2018	12	pendrive	17320
193	2015	1	RAM	12210
194	2015	2	RAM	15000
195	2015	3	RAM	12900
196	2015	4	RAM	14390
197	2015	5	RAM	9390
198	2015	6	RAM	8990
199	2015	7	RAM	7820
200	2015	8	RAM	8010
201	2015	9	RAM	6980
202	2015	10	RAM	7677
203	2015	11	RAM	8090
204	2015	12	RAM	6700
217	2016	1	RAM	9700
218	2016	2	RAM	14700
219	2016	3	RAM	7900
220	2016	4	RAM	12000
221	2016	5	RAM	14050
222	2016	6	RAM	15600
223	2016	7	RAM	12600
224	2016	8	RAM	12800
225	2016	9	RAM	17890
226	2016	10	RAM	13490
227	2016	11	RAM	12335
228	2016	12	RAM	13500
229	2017	1	RAM	11800
230	2017	2	RAM	17300
231	2017	3	RAM	15690
232	2017	4	RAM	5700

233	2017	5	RAM	9080
234	2017	6	RAM	11800
235	2017	7	RAM	12380
236	2017	8	RAM	13220
237	2017	9	RAM	5320
238	2017	10	RAM	7000
239	2017	11	RAM	4890
240	2017	12	RAM	6600
241	2018	1	RAM	5600
242	2018	2	RAM	8090
243	2018	3	RAM	8690
244	2018	4	RAM	3400
245	2018	5	RAM	2720
246	2018	6	RAM	5069
247	2018	7	RAM	7870
248	2018	8	RAM	9800
249	2018	9	RAM	6500
250	2018	10	RAM	8800
251	2018	11	RAM	10900
252	2018	12	RAM	12890

240 rows in set (0.001 sec)

1. Slicing Operation:

MariaDB [dwm]> select * from fact_sales where product='pendrive';

id	year	month	product	sales_qty
145	2015	1	pendrive	18700
146	2015	2	pendrive	12900
147	2015	3	pendrive	19920
148	2015	4	pendrive	10190
149	2015	5	pendrive	11900
150	2015	6	pendrive	11800
151	2015	7	pendrive	18300
152	2015	8	pendrive	15670
153	2015	9	pendrive	12800
154	2015	10	pendrive	11900
155	2015	11	pendrive	12000
156	2015	12	pendrive	9010
157	2016	1	pendrive	13440
158	2016	2	pendrive	7040
159	2016	3	pendrive	16010
160	2016	4	pendrive	14500
161	2016	5	pendrive	12500
162	2016	6	pendrive	12920
163	2016	7	pendrive	11095
164	2016	8	pendrive	13080
165	2016	9	pendrive	12900
166	2016	10	pendrive	11200
167	2016	11	pendrive	12800
168	2016	12	pendrive	17800
169	2017	1	pendrive	19800
170	2017	2	pendrive	11900

171	2017	3	pendrive	9500
172	2017	4	pendrive	15000
173	2017	5	pendrive	17020
174	2017	6	pendrive	10220
175	2017	7	pendrive	11230
176	2017	8	pendrive	18050
177	2017	9	pendrive	12350
178	2017	10	pendrive	12350
179	2017	11	pendrive	10220
180	2017	12	pendrive	17620
181	2018	1	pendrive	16000
182	2018	2	pendrive	18720
183	2018	3	pendrive	12902
184	2018	4	pendrive	15020
185	2018	5	pendrive	17000
186	2018	6	pendrive	13900
187	2018	7	pendrive	12330
188	2018	8	pendrive	17800
189	2018	9	pendrive	15400
190	2018	10	pendrive	11200
191	2018	11	pendrive	11320
192	2018	12	pendrive	17320

+-----+------+-------+----------+-----------+

48 rows in set (0.001 sec)

```
MariaDB [dwm]> select sum(sales_qty) from fact_sales where year=2015;
+----------------+
| sum(sales_qty) |
+----------------+
|         570314 |
+----------------+
1 row in set (0.001 sec)
```

2. Dice Operation:

```
MariaDB [dwm]> select * from fact_sales where product='pendrive' and year='2015';
+-----+------+-------+----------+-----------+
| id    | year | month | product    | sales_qty |
+-----+------+-------+----------+-----------+
| 145 | 2015 | 1         | pendrive |         18700 |
| 146 | 2015 | 2         | pendrive |         12900 |
| 147 | 2015 | 3         | pendrive |         19920 |
| 148 | 2015 | 4         | pendrive |         10190 |
| 149 | 2015 | 5         | pendrive |         11900 |
| 150 | 2015 | 6         | pendrive |         11800 |
| 151 | 2015 | 7         | pendrive |         18300 |
| 152 | 2015 | 8         | pendrive |         15670 |
| 153 | 2015 | 9         | pendrive |         12800 |
| 154 | 2015 | 10       | pendrive |         11900 |
| 155 | 2015 | 11       | pendrive |         12000 |
| 156 | 2015 | 12       | pendrive |           9010 |
+-----+------+-------+----------+-----------+
12 rows in set (0.001 sec)
```

3. Rollup Operation:

```
MariaDB [dwm]> select product, sum(sales_qty) from fact_sales group by product;
+-----------+----------------+
| product      | sum(sales_qty) |
+-----------+----------------+
| hard disk |              463922 |
| keyboard    |               377417 |
| monitor      |               351299 |
| pendrive    |               664547 |
| RAM           |               485831 |
+-----------+----------------+
5 rows in set (0.001 sec)

MariaDB [dwm]> select year, sum(sales_qty) from fact_sales group by year with rollup;
+------+----------------+
| year | sum(sales_qty) |
+------+----------------+
| 2015 |              570314 |
| 2016 |              601258 |
| 2017 |              620672 |
| 2018 |              550772 |
| NULL |             2343016 |
+------+----------------+
5 rows in set (0.001 sec)
```

4. Pivot Operation:

```
MariaDB [dwm]> select product,sum(if(year=2015,sales_qty,0))
    -> as '2015',sum(if(year=2016,sales_qty,0))
    -> as '2016',sum(if(year=2017,sales_qty,0))
    -> as '2017',sum(if(year=2018,sales_qty,0))
    -> as '2018' from fact_sales group by product;
+-----------+--------+--------+--------+--------+
| product   | 2015   | 2016   | 2017   | 2018   |
+-----------+--------+--------+--------+--------+
| hard disk | 136860 | 107620 | 120884 |  98558 |
| keyboard  |  79540 |  95690 | 114510 |  87677 |
| monitor   |  70667 |  86098 |  99238 |  95296 |
| pendrive  | 165090 | 155285 | 165260 | 178912 |
| RAM       | 118157 | 156565 | 120780 |  90329 |
+-----------+--------+--------+--------+--------+
5 rows in set (0.001 sec)
```

5. Drilldown Operation:

```
MariaDB [dwm]> select year,sum(if(month=1,sales_qty,0))
    -> as 'JAN',sum(if(month=2,sales_qty,0))
    -> as 'FEB',sum(if(month=3,sales_qty,0))
    -> as 'MAR',sum(if(month=4,sales_qty,0))
    -> as 'APR',sum(if(month=5,sales_qty,0))
    -> as 'MAY',sum(if(month=6,sales_qty,0))
    -> as 'JUNE',sum(if(month=7,sales_qty,0))
    -> as 'JUL',sum(if(month=8,sales_qty,0))
    -> as 'AUG',sum(if(month=9,sales_qty,0))
    -> as 'SEP',sum(if(month=10,sales_qty,0))
    -> as 'OCT',sum(if(month=11,sales_qty,0))
    -> as 'NOV',sum(if(month=12,sales_qty,0))
    -> as 'DEC' from fact_sales group by year;
+------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+
| year | JAN   | FEB   | MAR   | APR   | MAY   | JUNE  | JUL   | AUG   | SEP   | OCT   | NOV   | DEC   |
+------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+
| 2015 | 54149 | 48899 | 61000 | 34060 | 33850 | 43780 | 50390 | 51470 | 51540 | 46897 | 45699 | 48580 |
| 2016 | 34720 | 51850 | 39380 | 37900 | 57338 | 49306 | 56285 | 51030 | 50206 | 48668 | 66255 | 58320 |
| 2017 | 64082 | 61190 | 54840 | 45082 | 45040 | 48710 | 49890 | 55660 | 48270 | 50270 | 39258 | 58380 |
| 2018 | 40679 | 47398 | 42082 | 53030 | 40680 | 44108 | 44000 | 47660 | 41477 | 43490 | 47139 | 59029 |
+------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+
4 rows in set (0.003 sec)
```

Course Code: CSL603	Course Name: DWM LAB
Class: TE-CO	Batch: 3
Roll no: 18CO63	Name: SHAIKH TAUSEEF MUSHTAQUE ALI

Experiment :05

Aim: Implementation of Naive Bayes algorithm

Code:

```
# Naive Bayes


# Importing the libraries
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd


# Importing the dataset
dataset = pd.read_csv('Social_Network_Ads.csv')
X = dataset.iloc[:, [2, 3]].values
y = dataset.iloc[:, -1].values


# Splitting the dataset into the Training set and Test set
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.25, random_state = 0)


# Feature Scaling
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
X_train = sc.fit_transform(X_train)
X_test = sc.transform(X_test)
```

```
# Training the Naive Bayes model on the Training set

from sklearn.naive_bayes import GaussianNB
classifier = GaussianNB()
classifier.fit(X_train, y_train)

# Predicting the Test set results
y_pred = classifier.predict(X_test)

# Making the Confusion Matrix
from sklearn.metrics import confusion_matrix
cm = confusion_matrix(y_test, y_pred)
print(cm)

# Visualising the Training set results
from matplotlib.colors import ListedColormap
X_set, y_set = X_train, y_train
X1, X2 = np.meshgrid(np.arange(start = X_set[:, 0].min() - 1, stop = X_set[:, 0].max() + 1, step = 0.01),
                     np.arange(start = X_set[:, 1].min() - 1, stop = X_set[:, 1].max() + 1, step = 0.01))
plt.contourf(X1, X2, classifier.predict(np.array([X1.ravel(), X2.ravel()]).T).reshape(X1.shape),
             alpha = 0.75, cmap = ListedColormap(('red', 'green')))
plt.xlim(X1.min(), X1.max())
plt.ylim(X2.min(), X2.max())
for i, j in enumerate(np.unique(y_set)):
    plt.scatter(X_set[y_set == j, 0], X_set[y_set == j, 1],
                c = ListedColormap(('red', 'green'))(i), label = j)
plt.title('Naive Bayes (Training set)')
plt.xlabel('Age')
plt.ylabel('Estimated Salary')
plt.legend()
plt.show()
```

```
# Visualising the Test set results
from matplotlib.colors import ListedColormap
X_set, y_set = X_test, y_test
X1, X2 = np.meshgrid(np.arange(start = X_set[:, 0].min() - 1, stop = X_set[:, 0].max() + 1, step = 0.01),
                     np.arange(start = X_set[:, 1].min() - 1, stop = X_set[:, 1].max() + 1, step = 0.01))
plt.contourf(X1, X2, classifier.predict(np.array([X1.ravel(), X2.ravel()]).T).reshape(X1.shape),
             alpha = 0.75, cmap = ListedColormap(('red', 'green')))
plt.xlim(X1.min(), X1.max())
plt.ylim(X2.min(), X2.max())
for i, j in enumerate(np.unique(y_set)):
    plt.scatter(X_set[y_set == j, 0], X_set[y_set == j, 1],
                c = ListedColormap(('red', 'green'))(i), label = j)
plt.title('Naive Bayes (Test set)')
plt.xlabel('Age')
plt.ylabel('Estimated Salary')
plt.legend()
plt.show()
```

CSV FILE:
https://drive.google.com/file/d/1gIK6yLb3puhiLG13_83Dqw_7c3v_ntdn/view?usp=sharing

Output:

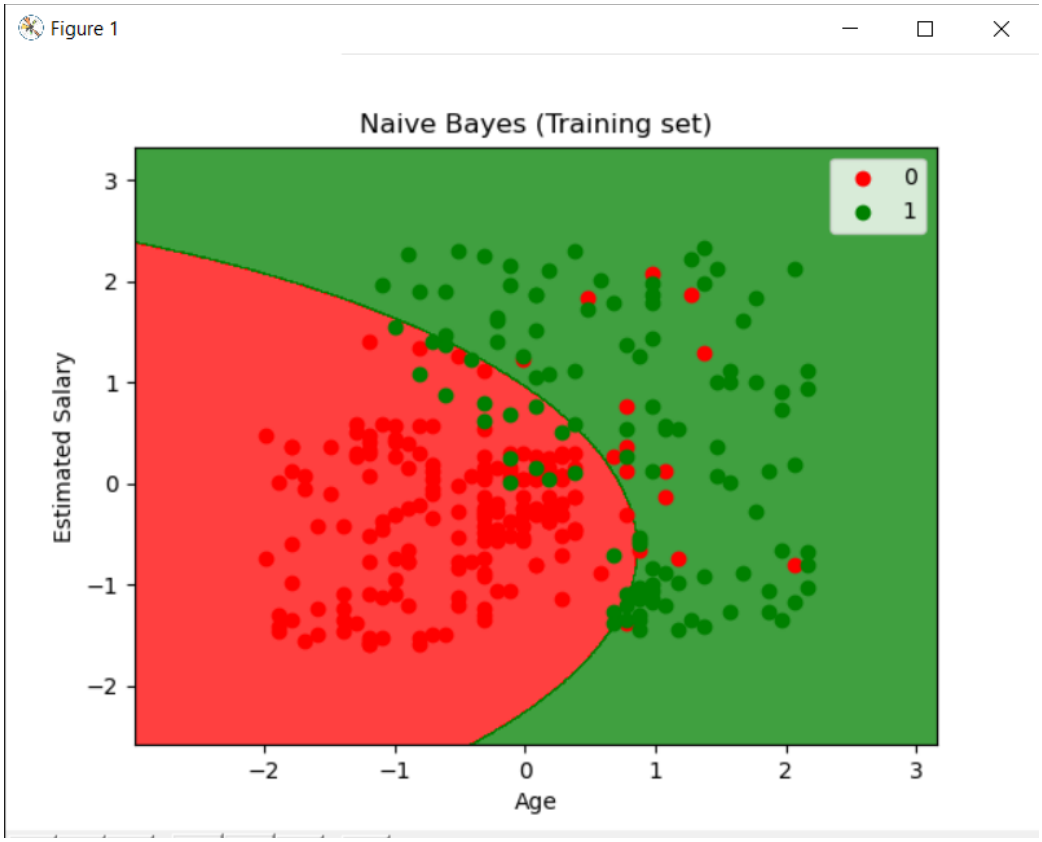


Figure 1
Naive Bayes (Training set)
Estimated Salary
Age
0
1
3
2
1
0
−1
−2
−2
−1
0
1
2
3

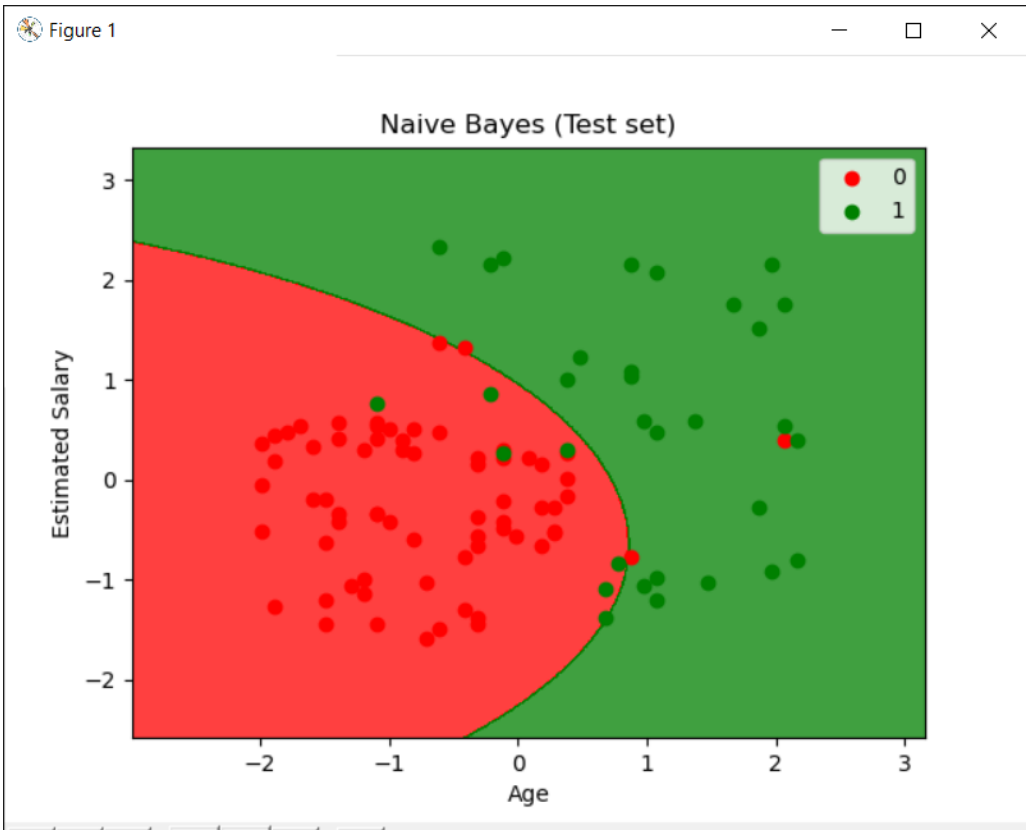


Figure 1
Naive Bayes (Test set)
Estimated Salary
Age
0
1
3
2
1
0
−1
−2
−2
−1
0
1
2
3

Course Code: CSL603	**Course Name: DWM LAB**
Class: TE-CO	**Batch: 3**
Roll no: 18CO63	**Name: SHAIKH TAUSEEF MUSHTAQUE ALI**

Experiment :06

Aim: Implementation of Decision Tree algorithm

Code:

```
# Decision Tree Classification


# Importing the libraries
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd


# Importing the dataset
dataset = pd.read_csv('Social_Network_Ads.csv')
X = dataset.iloc[:, [2, 3]].values
y = dataset.iloc[:, 4].values


# Splitting the dataset into the Training set and Test set
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.25, random_state = 0)


# Feature Scaling
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
X_train = sc.fit_transform(X_train)
X_test = sc.transform(X_test)
```

```
# Training the Decision Tree Classification model on the Training set
from sklearn.tree import DecisionTreeClassifier
classifier = DecisionTreeClassifier(criterion = 'entropy', random_state = 0)
classifier.fit(X_train, y_train)

# Predicting the Test set results
y_pred = classifier.predict(X_test)

# Making the Confusion Matrix
from sklearn.metrics import confusion_matrix
cm = confusion_matrix(y_test, y_pred)
print(cm)

# Visualising the Training set results
from matplotlib.colors import ListedColormap
X_set, y_set = X_train, y_train
X1, X2 = np.meshgrid(np.arange(start = X_set[:, 0].min() - 1, stop = X_set[:, 0].max() + 1, step = 0.01),
                     np.arange(start = X_set[:, 1].min() - 1, stop = X_set[:, 1].max() + 1, step = 0.01))
plt.contourf(X1, X2, classifier.predict(np.array([X1.ravel(), X2.ravel()]).T).reshape(X1.shape),
             alpha = 0.75, cmap = ListedColormap(('red', 'green')))
plt.xlim(X1.min(), X1.max())
plt.ylim(X2.min(), X2.max())
for i, j in enumerate(np.unique(y_set)):
    plt.scatter(X_set[y_set == j, 0], X_set[y_set == j, 1],
                c = ListedColormap(('red', 'green'))(i), label = j)
plt.title('Decision Tree Classification (Training set)')
plt.xlabel('Age')
plt.ylabel('Estimated Salary')
plt.legend()
plt.show()
```

```
# Visualising the Test set results
from matplotlib.colors import ListedColormap
X_set, y_set = X_test, y_test
X1, X2 = np.meshgrid(np.arange(start = X_set[:, 0].min() - 1, stop = X_set[:, 0].max() + 1, step = 0.01),
                     np.arange(start = X_set[:, 1].min() - 1, stop = X_set[:, 1].max() + 1, step = 0.01))
plt.contourf(X1, X2, classifier.predict(np.array([X1.ravel(), X2.ravel()]).T).reshape(X1.shape),
             alpha = 0.75, cmap = ListedColormap(('red', 'green')))
plt.xlim(X1.min(), X1.max())
plt.ylim(X2.min(), X2.max())
for i, j in enumerate(np.unique(y_set)):
    plt.scatter(X_set[y_set == j, 0], X_set[y_set == j, 1],
                c = ListedColormap(('red', 'green'))(i), label = j)
plt.title('Decision Tree Classification (Test set)')
plt.xlabel('Age')
plt.ylabel('Estimated Salary')
plt.legend()
plt.show()
```

CSV File:

https://drive.google.com/file/d/1gIK6yLb3puhiLG13_83Dqw_7c3v_ntdn/view

Output:

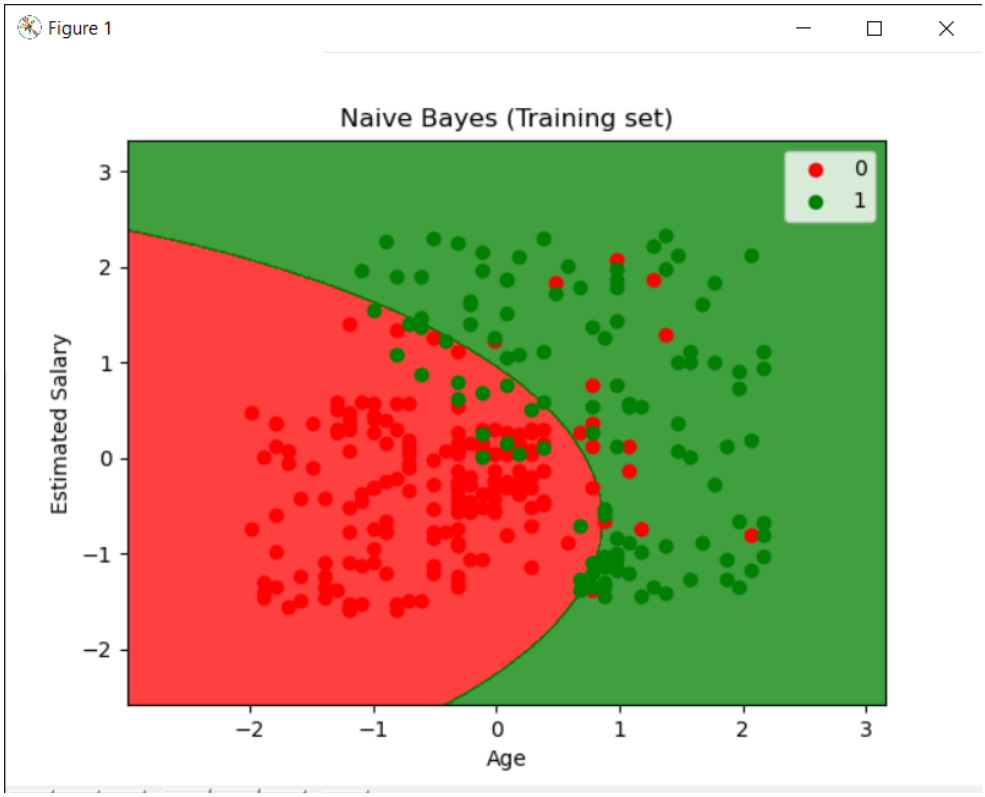


Figure 1
Naive Bayes (Training set)
Estimated Salary
Age
0
1
3
2
1
0
−1
−2
−2
−1
0
1
2
3

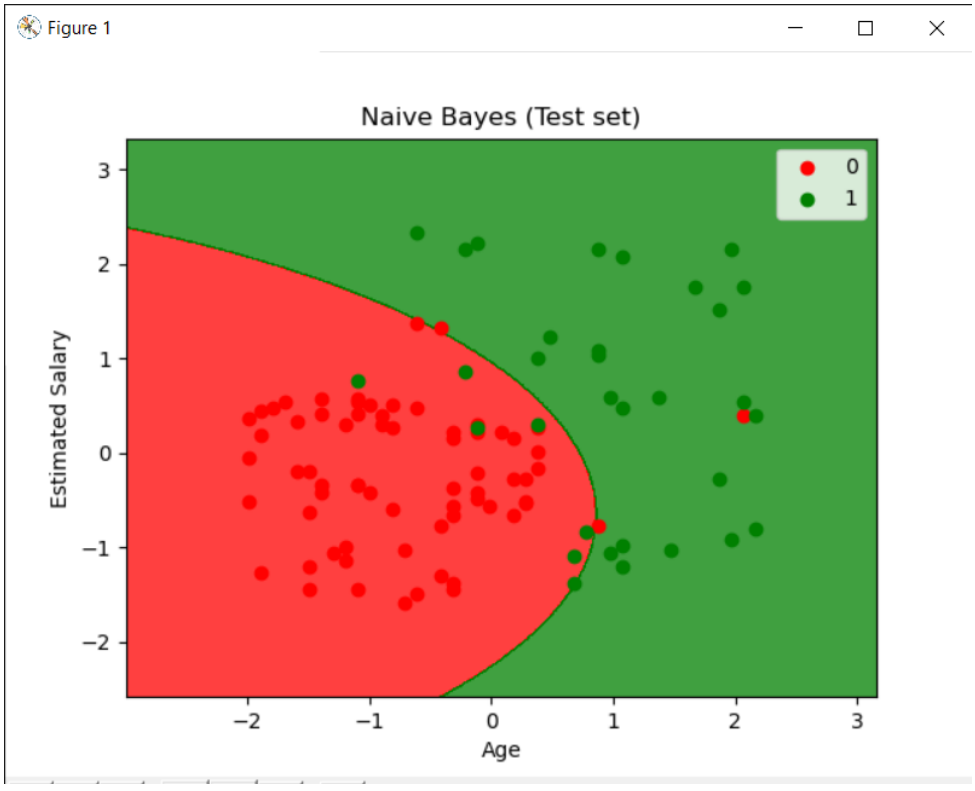


Figure 1
Naive Bayes (Test set)
Estimated Salary
Age
0
1
3
2
1
0
−1
−2
−2
−1
0
1
2
3

Course Code: CSL603	Course Name: DWM LAB
Class: TE-CO	Batch: 3
Roll no: 18CO63	Name: SHAIKH TAUSEEF MUSHTAQUE ALI

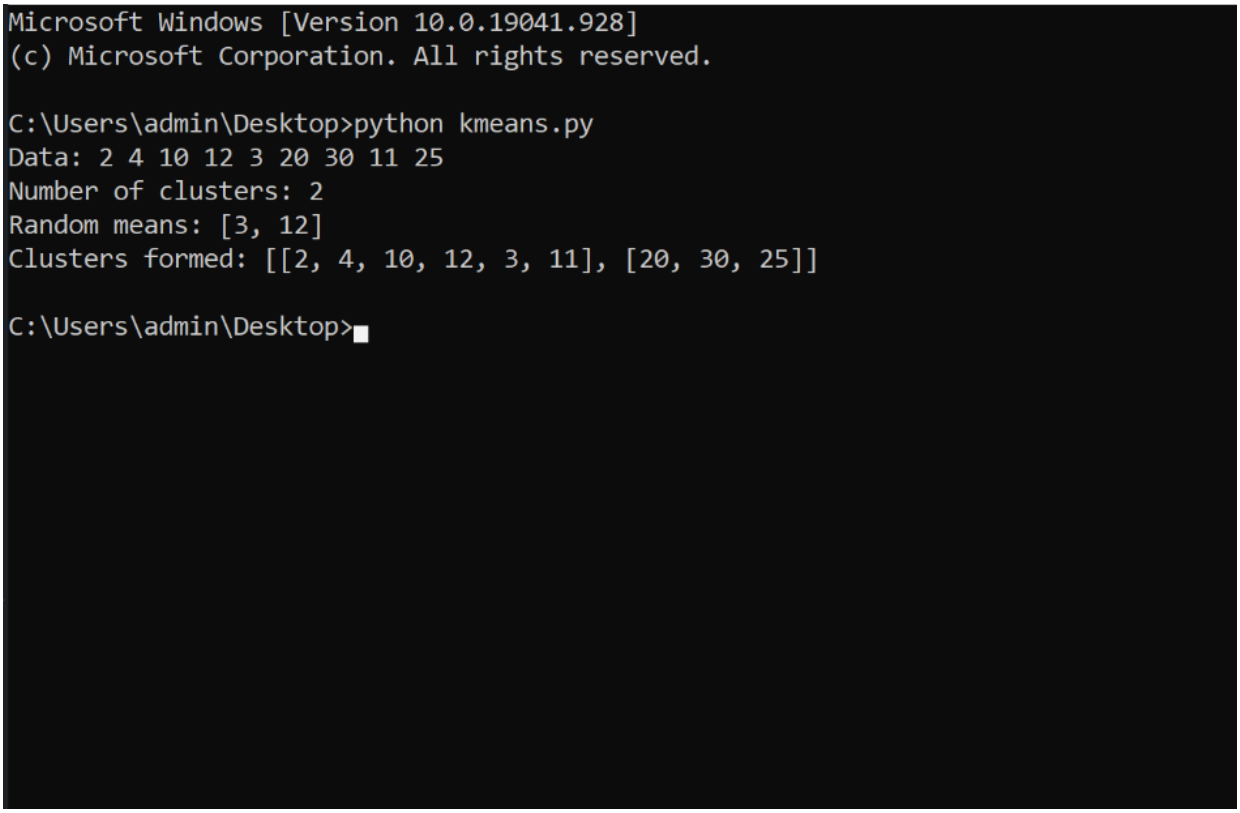
Experiment :07

Aim: Implementation of K-means algorithm

Code:

```
import random
arr = list(map(int, input("Data: ").split()))
k = int(input("Number of clusters: "))
m1 = random.choices(arr, k=k)
print("Random means:",m1)
m2 = []
K = []
for l in range(k):
    K.append([])
for i in range(len(arr)):
    n = 0
    mini = float("inf")
    p = 0
    for j in range(k):
        if abs(m1[j]-arr[i])<mini:
            mini = abs(m1[j]-arr[i])
            n = arr[i]
            p = j
    K[p].append(n)
while (m1!=m2):
    for l in range(k):
        m2.append(sum(K[l])/len(K[l]))
        K[l].clear()
    #print("Means:",m2)
    for i in range(len(arr)):
        n = 0
        mini = float("inf")
        p = 0
        for j in range(k):
            if abs(m2[j]-arr[i])<mini:
                mini = abs(m2[j]-arr[i])
                n = arr[i]
                p = j
        K[p].append(n)
    if m1!=m2:
        m1 = m2.copy()
        m2.clear()
    else:
        break
else:
    print("Clusters formed:",K)
print("Clusters formed:",K)
```

Output:


Microsoft Windows [Version 10.0.19041.928]
(c) Microsoft Corporation. All rights reserved.
C:\Users\admin\Desktop>python kmeans.py
Data: 2 4 10 12 3 20 30 11 25
Number of clusters: 2
Random means: [3, 12]
Clusters formed: [[2, 4, 10, 12, 3, 11], [20, 30, 25]]
C:\Users\admin\Desktop>

Course Code: CSL603	Course Name: DWM LAB
Class: TE-CO	Batch: 3
Roll no: 18CO63	Name: SHAIKH TAUSEEF MUSHTAQUE ALI

Experiment :08

Aim: Implementation of Apriori algorithm

Code:

```
import sys

from itertools import chain, combinations
from collections import defaultdict
from optparse import OptionParser


def subsets(arr):
    """ Returns non empty subsets of arr"""
    return chain(*[combinations(arr, i + 1) for i, a in enumerate(arr)])


def returnItemsWithMinSupport(itemSet, transactionList, minSupport, freqSet):
    """calculates the support for items in the itemSet and returns a subset
    of the itemSet each of whose elements satisfies the minimum support"""
    _itemSet = set()
    localSet = defaultdict(int)

    for item in itemSet:
        for transaction in transactionList:
            if item.issubset(transaction):
                freqSet[item] += 1
                localSet[item] += 1

    for item, count in localSet.items():
        support = float(count) / len(transactionList)

        if support >= minSupport:
            _itemSet.add(item)

    return _itemSet


def joinSet(itemSet, length):
    """Join a set with itself and returns the n-element itemsets"""
    return set(
        [i.union(j) for i in itemSet for j in itemSet if len(i.union(j)) == length]
    )


def getItemSetTransactionList(data_iterator):
    transactionList = list()
    itemSet = set()
```

```
    for record in data_iterator:
        transaction = frozenset(record)
        transactionList.append(transaction)
        for item in transaction:
            itemSet.add(frozenset([item]))    # Generate 1-itemSets
    return itemSet, transactionList


def runApriori(data_iter, minSupport, minConfidence):
    """
    run the apriori algorithm. data_iter is a record iterator
    Return both:
     - items (tuple, support)
     - rules ((pretuple, posttuple), confidence)
    """
    itemSet, transactionList = getItemSetTransactionList(data_iter)

    freqSet = defaultdict(int)
    largeSet = dict()
    # Global dictionary which stores (key=n-itemSets,value=support)
    # which satisfy minSupport

    assocRules = dict()
    # Dictionary which stores Association Rules

    oneCSet = returnItemsWithMinSupport(itemSet, transactionList, minSupport, freqSet)

    currentLSet = oneCSet
    k = 2
    while currentLSet != set([]):
        largeSet[k - 1] = currentLSet
        currentLSet = joinSet(currentLSet, k)
        currentCSet = returnItemsWithMinSupport(
            currentLSet, transactionList, minSupport, freqSet
        )
        currentLSet = currentCSet
        k = k + 1

    def getSupport(item):
        """local function which Returns the support of an item"""
        return float(freqSet[item]) / len(transactionList)

    toRetItems = []
    for key, value in largeSet.items():
        toRetItems.extend([(tuple(item), getSupport(item)) for item in value])

    toRetRules = []
    for key, value in list(largeSet.items())[1:]:
        for item in value:
            _subsets = map(frozenset, [x for x in subsets(item)])
            for element in _subsets:
                remain = item.difference(element)
                if len(remain) > 0:
                    confidence = getSupport(item) / getSupport(element)
                    if confidence >= minConfidence:
                        toRetRules.append(((tuple(element), tuple(remain)), confidence))
    return toRetItems, toRetRules
```

```
def printResults(items, rules):
    """prints the generated itemsets sorted by support and the confidence rules sorted by confidence"""
    for item, support in sorted(items, key=lambda x: x[1]):
        print("item: %s , %.3f" % (str(item), support))
    print("\n------------------------ RULES:")
    for rule, confidence in sorted(rules, key=lambda x: x[1]):
        pre, post = rule
        print("Rule: %s ==> %s , %.3f" % (str(pre), str(post), confidence))


def to_str_results(items, rules):
    """prints the generated itemsets sorted by support and the confidence rules sorted by confidence"""
    i, r = [], []
    for item, support in sorted(items, key=lambda x: x[1]):
        x = "item: %s , %.3f" % (str(item), support)
        i.append(x)

    for rule, confidence in sorted(rules, key=lambda x: x[1]):
        pre, post = rule
        x = "Rule: %s ==> %s , %.3f" % (str(pre), str(post), confidence)
        r.append(x)

    return i, r


def dataFromFile(fname):
    """Function which reads from the file and yields a generator"""
    with open(fname, "rU") as file_iter:
        for line in file_iter:
            line = line.strip().rstrip(",")  # Remove trailing comma
            record = frozenset(line.split(","))
            yield record


if __name__ == "__main__":

    optparser = OptionParser()
    optparser.add_option(
        "-f", "--inputFile", dest="input", help="filename containing csv", default=None
    )
    optparser.add_option(
        "-s",
        "--minSupport",
        dest="minS",
        help="minimum support value",
        default=0.15,
        type="float",
    )
    optparser.add_option(
        "-c",
        "--minConfidence",
        dest="minC",
        help="minimum confidence value",
        default=0.6,
        type="float",
    )

    (options, args) = optparser.parse_args()
```

```
inFile = None
if options.input is None:
        inFile = sys.stdin
elif options.input is not None:
        inFile = dataFromFile(options.input)
else:
        print("No dataset filename specified, system with exit\n")
        sys.exit("System will exit")

minSupport = options.minS
minConfidence = options.minC

items, rules = runApriori(inFile, minSupport, minConfidence)

printResults(items, rules)
```

Course Code: CSL603	**Course Name: DWM LAB**
Class: TE-CO	**Batch: 3**
Roll no: 18CO63	**Name: SHAIKH TAUSEEF MUSHTAQUE ALI**

Experiment :09

Aim: Case study of any one BI tool like Oracle BI, SPSS, Clementine, and XL Miner etc.

CASE STUDY:

P.T.O.

The Journal of Supercomputing (2018) 74:5171–5186
https://doi.org/10.1007/s11227-018-2479-2

A security review of local government using NIST CSF: a case study

Ahmed Ibrahim[1] · **Craig Valli**[1] · **Ian McAteer**[1] · **Junaid Chaudhry**[2]

Published online: 12 July 2018

Abstract
Evaluating cyber security risk is a challenging task regardless of an organisation's nature of business or size, however, an essential activity. This paper uses the National Institute of Standards and Technology (NIST) cyber security framework (CSF) to assess the cyber security posture of a local government organisation in Western Australia. Our approach enabled the quantification of risks for specific NIST CSF core functions and respective categories and allowed making recommendations to address the gaps discovered to attain the desired level of compliance. This has led the organisation to strategically target areas related to their people, processes, and technologies, thus mitigating current and future threats.

Keywords NIST cyber security framework · Local government · Cyber security · Risk assessment

1 Introduction

The National Institute of Standards and Technology (NIST) Cyber Security Framework (CSF) [28] is a risk-based approach to manage risks organisations face from a cyber security perspective. Similarly, several frameworks such as NIST SP 800-53 [27], COBIT5 [17], ISO/IEC 27001:2013 [23], ISA 62443-2-1:2009 [21], and ISA 62443-3-3:2013 [22] are being used to assess cyber security risk from different perspectives and outcomes are measured using different yardsticks. Often, navigating the various frameworks can be challenging for organisations, especially if such expertise are not present internally. Given the rapidly changing technology and threat landscape,

✉ Ahmed Ibrahim
ahmed.ibrahim@ecu.edu.au

1 Security Research Institute, School of Science, Edith Cowan University, 270 Joondalup Drive, Perth, WA 6027, Australia

2 College of Security and Intelligence, Embry-Riddle Aeronautical University, Prescott, AZ, USA

assessing the cyber security posture of an organisation, regardless of their business or size, is paramount.

Our focus of this paper is to demonstrate the application (Sect. 3) of NIST CSF in a local government organisation and provide recommendations (Sect. 5) based on our findings (Sect. 4).

The main contributions of this paper are:

- The adoption of the NIST CSF as an Assessment Tool and targeting different levels of the organisation, depending on their level of expertise and job function to obtain responses to facilitate assessment.
- Quantification of the assessment to reflect severity of actual risk, which in turn enabled the organisation to effectively address the issues to attain desired level of compliance.
- A detailed review of similar frameworks used in the industry and relevant case studies (Sect. 6).

The next section provides a background of the NIST CSF and its components. We recommend the reader to refer to NIST [28] for additional details and strategies for suitable approaches to implement, which would vary from organisation to organisation.

2 The NIST CSF

The NIST CSF [28] consists of the *Framework Core*, the *Framework Implementation Tiers*, and the *Framework Profiles*. The Framework Core consists of five concurrent and continuous functions; *Identify*, *Protect*, *Detect*, *Respond*, and *Recover*. We designed an Assessment Tool for our investigation based on these functions, which provided a systematic approach to ascertain the organisations cyber security risk management practices and processes.

The Framework Implementation Tiers describe the level an organisations cyber security risk management practices that comply with the framework. Tiers provide context and degree to which cyber security risks are managed and extent to which business needs are considered in cyber security risk management. The Assessment Tool enabled the determination of the organisations *Current Tier* based on various internal and external factors such as their risk management practices, threat environment, legal and regulatory requirements, business/mission objectives, and organisational constraints. Organisations should also determine the *Desired Tier*, provided it is feasible to implement, reduces cyber security risks, and meets the organisational goals. The following are descriptions of the tier levels [28]:

- *Tier-1 (Partial):* risk management practices are not formalised and managed in an ad hoc manner, lack awareness of cyber security risks organisation wide, and do not have processes in place to collaborate with external entities.
- *Tier-2 (Risk Informed):* risk management practices are formalised but not integrated organisation wide, but cyber security activities are prioritised based on risks with adequate means to perform related duties, with informal means to communicate cyber security information internally and externally.

- *Tier-3 (Repeatable):* risk management practices are formalised and policies are in place and are adaptable to cyber threats. Organisation-wide approach is required to manage cyber security with skilled and knowledgeable personnel to respond and understand dependencies and role of external partners.
- *Tier-4 (Adaptive):* cyber security practices are based on lessons learnt and predictive indicators, with continuous improvement, adaptability, and timely response. Organisation-wide approach to manage cyber security risks is part of the organisational culture and actively shares with external partners.

The Framework Profile represents the outcomes based on the business needs the organisation characterised from the Framework Core and determined using the Assessment Tool. Consequently, a *Current Profile* (the "as is" state) and a *Target Profile* (the "to be" state) can be used to identify opportunities for improving the cyber security of the organisation [28]. Framework profiles can be determined based on particular implementation scenarios, and therefore, the gap between Current Profile and Target Profile would vary as per scenario. In this paper, a local government-specific approach to CSF was adapted. However, industry-specific tailoring may be performed for the CSF.

3 Methodology

The NIST CSF allowed us to design an Assessment Tool targeted at three levels of participants within the organisation, i.e. executive, management and technical. The rationale was to ascertain organisation-wide understanding of cyber security risks. Hence, the Assessment Tool comprised of questions addressing the requirements outlined as per the NIST CSF.

The questions were selected based on the nature and relevance to the level of participant. This is because the NIST CSF comprised of questions that were both technical and non-technical. Therefore, it would have been unrealistic to expect deep knowledge of technical operations or implementation level details from a policy level executive.

In order to assist us determine a baseline (i.e. the Desired Tier), additional questions were included in the Assessment Tool to determine the nature of the organisation and its business. This was then followed by the remaining requirements comprised in the NIST CSF.

3.1 Determining compliance

The compliance for each measure was based on the responses provided by the participants. They were graded as either, *Complaint*, *Partially Compliant*, or *Non-Compliant*; and each was assigned scores of either 10, 5, or 0, respectively, for each core function's subcategory. Any subcategory that was not applicable depending on the Desired Tier level was excluded from the compliance score calculation.

Given the number of security requirements for each Core Function's subcategory is N, then the number of applicable requirements in each subcategory given the Desired

Tier level is N'. Therefore, the total compliance score C for each core function's category can be defined as:

$$C = \frac{\Sigma R}{\Sigma N' \times 10} \tag{1}$$

where R is the compliance score for each category of the respective Core Function.

Additionally, a detailed document audit was conducted on existing policies and procedures. The Information Technology (IT) infrastructure (internal, remote locations, and cloud) were reviewed, and a detailed internal vulnerability assessment was also conducted during our investigation.

4 Findings

The responses provided by the Executive, Management, and Technical participants gave insight into the organisation's cyber security posture. Table 1 shows the summary of the compliance of NIST CSF assessment. The compliance scores were determined based on Eq. 1 presented previously.

For Identify core function, the organisation scored 36%. Their ability to track assets centrally, keep management informed, and understand operational risks from a cyber security perspective was limited, while a strategy to manage such risks did not exist. However, the organisation understood its business well and were able set priorities to support risk management decisions.

Access to physical/virtual assets were through authorisation and well-defined processes. The staff were trained and informed adequately of information security related duties and responsibilities. Certain aspects of data security related to confidentiality and availability were done reasonably well, however, assuring integrity of data needed improvement. Similarly, local maintenance and remote maintenance of IT infrastructure were carried out in a manner consistent to policies and procedures. However, relevant policies, processes, and procedures, as well as technology to assist the protection of information systems and relevant assets, were lacking. Therefore, in aggregate, the organisation scored 45% compliance for Protect core function.

The organisation scored weakest in the detection of cyber security incidents with a score of 25%. Although certain monitoring activities were in place to track physical security and malicious code, timely detection of anomalous activities and detection processes were lacking or non-existent.

Despite the lack of a specific response plan to respond to a cyber security events, the organisation had measures in place to report incidents and coordinate activities to respond adequately, which resulted in a 38% compliance score for Respond core function. These practices are updated from time to time; however, mechanism to perform post-incident analysis or to mitigate future cyber security events has not been implemented presently.

Interestingly, the organisation was well prepared to deal with recovery and resumption of core services after a cyber security event. The recovery plans in place are tested, updated, and improved periodically, thus receiving full compliance for Recover core functionality of the framework.

Table 1 NIST CSF compliance matrix

Function	Category	Compliance (%)	Total (%)
Identify (ID)	Asset Management (ID.AM)	33	36
	Business Environment (ID.BE)	75	
	Governance (ID.GV)	25	
	Risk Assessment (ID.RA)	25	
	Risk Management Strategy (ID.RM)	0	
Protect (PR)	Access Control (PR.AC)	60	45
	Awareness and Training (PR.AT)	70	
	Data Security (PR.DS)	50	
	Information Protection Processes and Procedures (PR.IP)	20	
	Maintenance (PR.MA)	75	
	Protective Technology (PR.PT)	38	
Detect (DE)	Anomalies and Events (DE.AE)	0	25
	Security Continuous Monitoring (DE.CM)	43	
	Detection Processes (DE.DP)	25	
Respond (RS)	Response Planning (RS.RP)	0	38
	Communications (RS.CO)	88	
	Analysis (RS.AN)	0	
	Mitigation (RS.MI)	0	
	Improvements (RS.IM)	100	
Recover (RC)	Recovery Planning (RC.RP)	100	100
	Improvements (RC.IM)	100	
	Communications (RC.CO)	100	

5 Recommendations

Based on the findings, the following recommendations were made with respect to each core function of the NIST CSF.

5.1 Identify

(a) Establish a central inventory of assets, including physical devices and systems, software, and external systems with all required information and prioritise based on classification, criticality, and business value.
(b) Identify the organisations role in the supply chain (i.e. producer-consumer model) as it captures and retains public data, collects revenue, and provides services to its stakeholders.
(c) Establish an Information Security policy and reference relevant federal and state policies regarding cyber security to ensure legal and regulatory requirements are understood and managed.

(d) Identify and prioritise threats and vulnerabilities, both internal and external, to determine cyber security risks to the organisations operations, assets, and individuals.
(e) Establish risk management processes that are managed and agreed to by stakeholders to support operational risk decisions.

5.2 Protect

(a) Strengthen the Access Control policy and procedures for organisation-wide assets that require both physical and remote access.
(b) Sensitise and increase awareness about cyber security throughout the workforce more comprehensively and provide adequate cyber security training based on roles and responsibilities. In this regard, clearly describe cyber security roles and responsibilities for relevant staff and external stakeholders.
(c) Enforce required provisions for data security in the policy and implement data-at-rest and data-in-transit security, and integrity-checking mechanisms to ensure confidentiality, integrity, and availability of information and data.
(d) Establish required policies, processes, and procedures to manage protection of information assets. This include establishment of lacking policies and processes, particularly for configuration management, data destruction, and physical operating environment; identification of security baselines; SDLC for system management; formulate vulnerability, response, and recovery plans.
(e) Strengthen processes that control and log remote access to organisational assets by external maintenance contractors.
(f) Establish a central log of organisation-wide information systems and devices, establish Removable Media policy, and strengthen network segregation to protect communications and controls networks.

5.3 Detect

(a) Determine baselines for network operations and data flows, implement appropriate activities to detect and analyse events based on event data aggregated from multiple sources and sensors. Determine incident impact and threshold to prepare and allocate resources appropriately.
(b) Implement tools to monitor cyber and physical environments to detect unauthorised mobile code, external service provider activities, and unauthorised access. Perform organisation-wide vulnerabilities regularly.
(c) Outline detection requirements in Information Security policy and continuously improve these processes to ensure timely and adequate awareness of anomalous events.

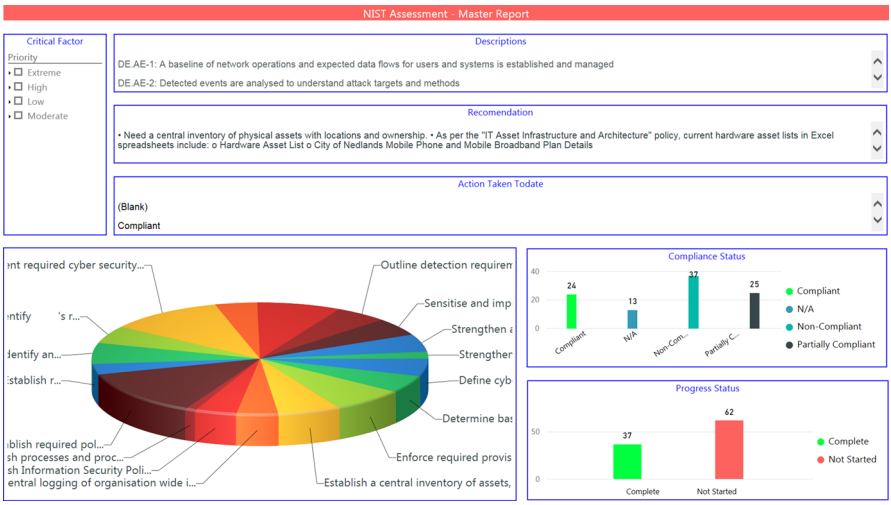


Fig. 1 Microsoft Power BI Internal Site for tracking, visualising, and reporting NIST CSF assessment findings, courtesy of the participating local government organisation

5.4 Respond

(a) Establish processes and procedures to respond to cyber security events in a timely manner.
(b) Define cyber security roles and responsibilities in Information Security policy to ensure activities are coordinated for internal and external stakeholders including law enforcement in response to cyber security events.
(c) Implement required cyber security events notification and detection systems to ensure adequate information is available to analyse and understand the impact to support recovery activities.
(d) Implement required cyber security controls to detect, report, and contain incidents to prevent escalation of an incident, mitigate its effect, and eradicate the incidents.

Each of the above recommendations also had specific internal stakeholder(s) identified to indicate ownership and responsibility for addressing the issues associated. Consequently, the organisation was then able to develop strategies to address the issues identified, and assign specific tasks to individuals. For this purpose, the organisation established an internal document using Microsoft Power BI [25] (typically referred to as a Power BI site) to track and visualise the status of the NIST CSF assessment (Fig. 1).

The Power BI site facilitated transparency, visibility, and central reporting throughout the organisation. Intuitively, this resulted in a rapid and responsive drive for the organisation to address and prioritise issues based on severity and cost, with the goal of achieving Tier-2 compliance.

Furthermore, a desire to achieve a higher compliance level such as Tier-3 was expressed. Such aspiration is encouraged, however, with caution. Even though a higher level of compliance will improve the cyber security posture of the organisation, it will

also affect other aspects such as resources and cost. For example, when contrasting the Risk Management Process between Tier-2 and Tier-3 as defined in the NIST CSF [28]:

(a) Implementation of risk management practices are not mandatory in Tier-2, whereas these have to be implemented as organisation-wide policies in Tier-3. Thus, Tier-3 organisations should have procedures, processes, technology, and human resources to implement relevant policies.
(b) The cyber security activities' priorities are updated in a passive nature in Tier-2 as opposed to regular active updates and constant re-evaluation of priorities for Tier-3 compliance. To acquire such capability, an organisation requires adequate technology, skilled human resources, and relevant policies that would enable keeping pace with the changes in the technology and threat landscape.

In addition to the two points highlighted above, considering both Integrated Risk Management Program and External Participation [28], significant investment in resources and human skills development or acquisition is needed to make the transition from Tier-2 to Tier-3. Moreover, this should only be considered carefully based on the organisation's business requirements, strategic objectives, budget, risk appetite, and current and future threats.

6 Related frameworks

The diversity and complexity of Information Technology (IT) system components have increased significantly in recent years. Consequently, in order for businesses to adequately secure these systems, several standards and frameworks have been developed [2]. Such frameworks need to be applicable to all manner of business sectors, be they government or private, enterprise or small-business. Tables 2 and 3 provide a summary of useful examples of how both NIST SP 800-53 and ISO/IEC 27001:2013 frameworks have been applied in practice.

Since NIST CSF can be considered as a high-level abstraction of related frameworks, it provides references to other related frameworks for specific implementation guidelines. These referenced frameworks include:

– NIST SP 800-53 Rev. 4.
– Control Objectives for Information and Related Technologies (COBIT5).
– ISO/IEC 27001:2013.
– ISA 62443-2-1:2009.
– ISA 62443-3-3:2013.

These are further described below.

6.1 NIST SP 800-53 Rev. 4

NIST SP 800-53 [27] revisions are made according to changes in responsibilities under the Federal Information Security Management Act (FISMA), Public Law (P.L.) 107-347. The latest version of this framework consists of five functions (Identify,

Table 2 Summary of case studies for NIST SP 800-53

Case study	Description
Maroochy water services cyber attack against critical infrastructure in 2000 [1]	Disgruntled former employee used insider knowledge, stolen configuration and equipments to release more than one million litres of untreated sewerage water resulting in considerable environmental damage and prosecution by the Environmental Protection Agency. The case study revealed that the application of CSF controls would have mitigated the cyber attack
Intel's high-risk IT business units' pilot project [11]	Intel IT's Office and Enterprise business units, considered to be high-level risk environments, were the testbed for a pilot project to test the effectiveness of the NIST SP 800-53. The benefits of using this framework were realised within a short timeframe, with coherent use of risk management technologies across the business model, improved identification of strengths and weaknesses, and more efficient assessments of security priorities. As a result of the pilot project, Intel IT planned to expand the framework's implementation throughout their business infrastructure
Cyber security framework implementation at the University of Chicago [32]	The University of Chicago used the framework to establish cyber security protection for Biological Services Division (BSD). The four-part implementation consisted of identifying the initial state of cyber security processes, assessment of the initial threat landscape, determination of the desired target status, and a roadmap to establish and monitor progress. This resulted in better identification of security requirements and target objectives, develop and maintain departmental processes to achieve these objectives, provide long-term security solutions in a cost-effective manner, and improve information-sharing and good work practices across departments with different cyber security requirements
How the University of Pittsburgh is using the NIST cyber security framework [31]	The University of Pittsburgh used the NIST SP 800-53 to implement an IT security package that would cater for diversified needs while enabling collaboration between different departments, accommodate a wide variety of information types and sensitivities, and encompass third-party contractors on an ad hoc basis. NIST SP 800-53 enabled these goals to be met through the streamlining of existing practices and improving documentation. The scalable nature of NIST CSF was applicable to the differing scope and IT requirements of each department within the University
SIEM-based framework for security controls automation [26]	The potential of using SIEM technology is investigated with the aim of maximising security-control automation. For the security controls identified in NIST SP 800-53, approximately 30% of these controls were considered as having the capability of automation control. The cost of implementing a SIEM-based framework for security-control automation would be quickly recouped within a short time compared to the reduced employee-hours required to monitor an infrastructure the size of a local government organisation

Table 2 continued

Case study	Description
Recommendations for information security awareness training for college students [24]	A survey largely based on NIST SP 800-50 was designed to assess information security awareness amongst students at the business college of a mid-sized University in New England. The survey found that less than one-quarter of the participants had undertaken any form of Information Security Awareness Training (ISAT), and only two of the 68 had enrolled in University-provided training. ISAT of employees in local government is an integral part of a well-implemented cyber security infrastructure. Any cyber security review needs to ascertain current levels of information security awareness to gauge whether existing training is effective or deficient. The training needs to be regularly updated as new vulnerabilities and threats continually develop in this field

Protect, Detect, Respond, and Recover), 22 categories, and 98 subcategories. This framework utilises a four-tier security model (Partial, Risk Informed, Repeatable, and Adaptive) and a seven-step process (Prioritise and Scope, Orient, Create a Current Profile, Conduct a Risk Assessment, Create a Target Profile, Determine, Analyse and Prioritise Gaps, and Implement Action Plan). It focuses on assessing the current situation by determining how to assess security, how to consider risk, and how to resolve the security threats.

6.2 Control Objectives for Information and Related Technologies (COBIT5)

COBIT5 [17] is a business CSF designed for the governance and maintenance of enterprise IT systems. It consists of five domains and 37 processes in line with the responsibility areas of plan, build, run, and monitor. COBIT5 is aligned and coordinated with other recognised IT standards and good practices, such as NIST, ISO 27000, COSO, ITIL, BiSL, CMMI, TOGAF and PMBOK. It is built around the following considerations:

- The need to meet stakeholder expectations.
- The end-to-end process control of the enterprise.
- To work as a single integrated framework.
- Recognising that "Management" and "Governance" are two different things.

6.3 ISO/IEC 27001:2013

ISO/IEC 27001:2013 [23] is an international information security standard published by the International Organisation for Standardisation (ISO) and the International Electrotechnical Commission (IEC), which originated from the British Standard, BS 7799. This framework consists of 114 controls in 14 groups describing the requirements needed to design and implement an Information Security Management Systems (ISMS). Version 2 released in 2013 replaces the 2005 version 1 edition. It is a standard

Table 3 Summary of case studies for ISO 27001:2013

Case study	Description
Thames Security Shredding (TSS) Ltd. [3]	TSS specialises in the collection and destruction of confidential documentation on a commercial scale. Maintaining information security is, therefore, a critical process to protect their clients' identity and to ensure compliance with the UK Data Protection Act 1998. Certification to the ISO/IEC 27001 standard was seen as an integral part of the implementation and maintenance of world-class customer-centric security controls that would satisfy existing and prospective customers' needs and allow for rapid growth in the business. ISO/IEC 27001 certification resulted in an improved attitude and awareness of their staff towards information-security-related issues. A risk-based business continuity plan was used to minimise the impact of any potential security breaches. The certification allowed documentation to be continually updated and improved as corrective actions were taken
Fredrickson International [4]	Debt collection is a sector which, like banking, finance, telecommunications, and local government, is coming under increasing scrutiny and regulation. Fredrickson International is a debt collection agency who lists a central government department, and several UK financial institutions and FTSE 100 companies amongst their clients. Since attaining ISO/IEC 27001 certification, Fredrickson has achieved higher levels of security awareness throughout its departments, staff, and employees. Security audits have become more streamlined, and customers were given the confidence that Fredrickson was conducting international best practice when it came to information security
Legal Ombudsman [5]	The office of the Legal Ombudsman for England and Wales was established to simplify the process by which members of the public, small businesses, charities, clubs, trusts, etc., could resolve complaints about legal practitioners. To improve its customer service, information security practices conducted by the office needed to show that greater information security awareness had been established, diligence and compliance in handling sensitive information were in place, and that an assurance framework was aligned with global best practice. ISO/IEC 27001 certification helped the Legal Ombudsman for England and Wales to provide clients with the confidence and reassurance that it was conducting its work by the highest work standards. A better understanding of the information security among its staff led to a reduction in risk and an increase in productivity
SVM Cards Europe [7]	SVM is a leading provider of gift card, voucher, e-code, reward code, and similar promotional and benefit schemes throughout Europe. SVM required secure business processes, improved internal organisation, increased information protection, and sought greater tender and competitive advantage. With ISO/IEC 27001 certification, SVM observed that processes became more of a lifestyle than strictly about security only, which resulted in less downtime, instigated a stronger organisational structure, improved on its ability to win new contracts, and have greater confidence that their information security processes were working properly

Table 3 continued

Case study	Description
InfoView Technologies [6]	InfoView Technologies, a Queensland-based data analytics company, required a business model that met state government requirements, improved data security understanding, became more professional, improved its business culture, and be able to sustain and continuously improve its information security management, systems, and policies. These goals were achieved through ISO/IEC 27001 accreditation, after which InfoView Technologies were able to gain increased market access, meet compliance requirements of the Queensland state government, reduce risk, become more competitive, and streamline its practices and business culture
Capgemini [8]	Capgemini is the largest IT services company in Europe; and a global leader in its multiple domains of services. Operating in more than 40 countries, and over 100 languages, Capgemini's business model needed to transcend national and cultural boundaries. Systems were required to be robust to avoid losing business and maintain competitiveness. Protection of client assets and resources was deemed a priority to assure confidentiality, integrity, and availability. Through ISO/IEC 27001 certification, Capgemini was able to ensure improved security within its departments and for its clients, enhance security awareness in its staff and employees, and provide more efficient and streamlined documentation and reporting procedures. Standards certification needed to be applicable within the global marketplace, and remain pertinent regardless of cultural differences
Costain [9]	Costain, a UK-based engineering and construction group, has contributed to the construction of significant projects worldwide. Obtaining standards certifications was seen as the correct path to achieve improvements in several internal processes. Such goals required the implementation of several standards, such as quality management standard (ISO 9001), environmental management (ISO 14001), health and safety (BS OHSAS 18001), collaborative business relationships (BS 11000), information security management (ISO/IEC 27001) and business continuity management (ISO 22301). Through the enactment of multiple standards, Costain was able to improve several areas of their business to the benefit of their internal and external customers

that should be instigated by all businesses where information security is a critical factor, but in particular, applies to software development, managed service providers/hosting services providers, IT, banking and insurance, information management, government agencies and their service providers, and E-commerce merchants [23].

6.4 ISA 62443-2-1:2009

ISA 62443-2-1:2009 [21] is an International Standards on Auditing (ISA) standard covering the elements required to develop an Industrial Automation Control System Security Management System (IACS-SMS). It consists of three categories, three ele-

ment groups, and 22 elements. The framework is the first of four ISA policy and procedure products that identifies the essentials necessary to establish an effective cyber security management system (CSMS). However, the step-by-step approach as to how this is achieved is company-specific and according to their own business culture. These essentials are:

- Risk analysis.
- Addressing risk with the CSMS.
- Monitoring and improving the CSMS.

6.5 ISA 62443-3-3:2013

ISA 62443-3-3:2013 [22] is an International Standards on Auditing (ISA) standard covering the elements required for cyber security controls of industrial control systems (ICS). It consists of seven Foundation Requirements and 51 System Requirements.

ISA 62443-3-3:2013 is the third of three ISA systems products, that outlines system security requirements and security levels [22].

6.6 Other frameworks

In addition to the above, other frameworks used in the industry include:

- *Committee of Sponsoring Organizations of the Treadway Commission (COSO)* is an enterprise risk management standard, designed jointly by five leading associations, with the aim of integrating strategy and performance [13].
- *Council on CyberSecurity Top 20 Critical Security Controls (CCS CSC)* consists of a prioritised set of actions, originally developed by the SANS Institute, to protect assets from cyber attack [12].
- *ISF Standard of Good Practice (SoGP)* is a standard aimed at providing controls and guidance on all aspects of information security [20].
- *ETSI Cyber Security Technical Committee (TC Cyber)* was developed to improve standards within the European telecommunications sector [15].
- *Sherwood Applied Business Security Architecture (SABSA) Enhanced NIST Cybersecurity (SENC)* project enhances the five core levels of the NIST CSF into a SABSA model consisting of a six-level security architecture [30].
- *IASME Consortium (IASME)* is an information assurance standard based on ISO 27000, but aimed at small businesses [18].
- *RFC 2196 - Site Security Handbook (SSH)* represents a guide on how to develop computer security policies and procedures [19].
- *Health Information Trust Alliance (HITRUST)* is the first IT security CSF designed specifically for the healthcare sector. It is based on existing NIST standards and is aimed at healthcare and information security professionals [16].
- *North American Electric Reliability Corporation Critical Infrastructure Protection (NERC-CIP) version 5* is a set of requirements needed to secure the assets of the North American bulk electric system [14].

– *Open Security Architecture (OSA)* is a free community-owned resource of advice on the selection, design, and integration of devices required to provide security and control of an IT network [29].
– *Good Practice Guide 13 (GPG13)* is a UK government CSF related to Code of Connection (CoCo) compliance for businesses to secure IT systems [10].

7 Conclusion

In this paper we have used the NIST CSF to evaluate the cyber security risks of a local government organisation in Western Australia. Our approach can be used to derive measurable metrics for each Framework Core function and respective categories, thus enabling the organisation to ascertain the cyber security preparedness to actual risk.

Our findings suggest that evaluating the Desired Tier compliance to the NIST CSF helps identify the specific people, process, and technology areas that require improvement (i.e. gaps), which directly influence threat mitigation. The application of CSF helped us understand the current security context of the organisation while identifying the risks and future growth areas to improve. While higher tier compliance maybe desired, we have also recommended that the organisation's business requirements, strategic goals, budget, risk appetite, and current and future threats to be considered carefully.

Furthermore, as we have presented several related frameworks, navigating such frameworks for self assessment can be challenging, often not intended by design even, but not impossible. We have observed that the NIST CSF offers an advantage over other frameworks in this regard. However, there is still room for developing additional tools that would simplify the implementation process and speed up adoption.

Therefore, our future work will aim to improve the current Assessment Tool we have used, with a focus of making it adaptable and accessible to a wider audience and measurable for accurate quantification of cyber preparedness.

Acknowledgements We would like to thank the Western Australia local government organisation for sharing their case study for this research. We would also like to thank their staff for their support and cooperation during the assessment.

References

1. Abrams M, Weiss J (2008) Malicious control system cyber security attack case study: Maroochy water services, Australia. https://www.mitre.org/sites/default/files/pdf/08_1145.pdf. Accessed 29 Jan 2018
2. Angelini M, Lenti S, Santucci G (2017) Crumbs: a cyber security framework browser. In: 2017 IEEE Symposium on Visualization for Cyber Security (VizSec). IEEE, pp 1–8

3. BSI Group (2011) Case study Thames Security Shredding (TSS) Ltd. https://www.bsigroup.com/Documents/iso-27001/case-studies/BSI-ISO-IEC-27001-case-study-Thames-Security-UK-EN.pdf?epslanguage=en-MY. Accessed 15 Feb 2018
4. BSI Group (2012) How Fredrickson has reduced third party scrutiny and protected its reputation with ISO 27001 certification. https://www.bsigroup.com/Documents/iso-27001/case-studies/BSI-ISO-IEC-27001-case-study-Fredrickson-International-EN-UK.pdf?epslanguage=en-MY. Accessed 15 Feb 2018
5. BSI Group (2013) Implementing best practice and improving client confidence with ISO/IEC 27001. https://www.bsigroup.com/Documents/iso-27001/case-studies/BSI-ISO-IEC-27001-case-study-Legal-Ombudsman-UK-EN.pdf. Accessed 15 Feb 2018
6. BSI Group (2013) Infoview case study. https://www.bsigroup.com/LocalFiles/EN-AU/_Case%20Studies/BSI%20Infoview%20Case%20Study.pdf. Accessed 15 Feb 2018
7. BSI Group (2013) Supporting business growth with ISO/IEC 27001. https://www.bsigroup.com/Documents/iso-27001/case-studies/BSI-ISO-IEC-27001-case-study-SVM-UK-EN.pdf. Accessed 15 Feb 2018
8. BSI Group (2014) Using ISO/IEC 27001 certification to increase resilience, reassure clients and gain a competitive edge. https://www.bsigroup.com/Documents/iso-27001/case-studies/BSI-ISO-IEC-27001-case-study-Capgemini-UK-EN.pdf. Accessed 15 Feb 2018
9. BSI Group (2015) Integrating management systems to improve business performance and achieve sustained competitive advantage. https://www.bsigroup.com/Documents/iso-22301/case-studies/Costain-case-study-UK-EN.pdf. Accessed 15 Feb 2018
10. Cabinet Office (2010) Gpg13: Protective monitoring controls. http://gpg13.com/executive-summary/. Accessed 13 Mar 2018
11. Casey T, Fiftal K, Landfield K, Miller J, Morgan D, Willis B (2015) The cybersecurity framework in action: an Intel use case. Intel Corporation, pp 1–10. https://supplier.intel.com/static/governance/documents/The-cybersecurity-framework-in-action-an-intel-use-case-brief.pdf. Accessed 30 Jan 2018
12. Center for Internet Security (2018) CIS controls. https://www.cisecurity.org/controls/. Accessed 6 Mar 2018
13. COSO (2017) Guidance on enterprise risk management. https://www.coso.org/Pages/erm.aspx. Accessed 6 Mar 2018
14. Elkins V (2014) Summary of CIP version 5 standards. http://www.velaw.com/uploadedfiles/vesite/resources/summarycipversion5standards2014.pdf. Accessed 12 Feb 2018
15. ETSI (2017) Overview of cybersecurity. https://www.enisa.europa.eu/events/enisa-cscg-2017/presentations/brookson. Accessed 7 Mar 2018
16. HITRUST (2017) Introduction to the HITRUST CSF. https://hitrustalliance.net/documents/csf_rmf_related/v9/CSFv9Introduction.pdf. Accessed 21 Mar 2018
17. IASCA (2012) Cobit 5. https://cobitonline.isaca.org/. Accessed 01 Feb 2018
18. IASME Consortium (2014) About cyber essentials. https://www.iasme.co.uk/cyberessentials/about-cyber-essentials/. Accessed 07 Mar 2018
19. IETF (1997) Rfc 2196: site security handbook. https://www.ietf.org/rfc/rfc2196.txt. Accessed 8 Mar 2018
20. Information Security Forum (2016) The ISF standard of good practice for information security. https://www.securityforum.org/tool/the-isf-standardrmation-security/. Accessed 8 Mar 2018
21. ISA (2009) ANSI/ISA-99.02.01-2009. http://www.icsdefender.ir/files/scadadefender-ir/paygahdanesh/standards/ISA-62443-2-1-Public.pdf. Accessed 13 Mar 2018
22. ISA (2012) ANSI/ISA-62443-3-3 (99.03.03)-2013. http://www.icsdefender.ir/files/scadadefender-ir/paygahdanesh/standards/ISA-62443-3-3-Public.pdf. Accessed 13 Mar 2018
23. ISO (2013) ISO/IEC 27001:2013. https://www.iso.org/standard/54534.html. Accessed 1 Feb 2018
24. Kim EB (2014) Recommendations for information security awareness training for college students. Inf Manag Comput Secur 22(1):115–126. https://doi.org/10.1108/IMCS-01-2013-0005
25. Microsoft (2018) Power BI. https://powerbi.microsoft.com/en-us/. Accessed 12 Apr 2018
26. Montesino R, Fenz S, Baluja W (2012) Siem-based framework for security controls automation. Inf Manag Comput Secur 20(4):248–263. https://doi.org/10.1108/09685221211267639
27. NIST (2014) Assessing security and privacy controls in federal information systems and organizations. http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-53Ar4.pdf. Accessed 1 Feb 2018

28. NIST (2014) Framework for improving critical infrastructure cybersecurity: Version 1.0. https://www.nist.gov/sites/default/files/documents/cyberframework/cybersecurity-framework-021214.pdf. Accessed 30 Jan 2018
29. OSA (2007) Osa landscape. http://www.opensecurityarchitecture.org/cms/foundations/osa-landscape. Accessed 15 Mar 2018
30. SABSA (2015) Project charter for the development of a SABSA enhanced nist cybersecurity framework. https://sabsa.org/sabsa-nist-framework-project/. Accessed 21 Mar 2018
31. Sweeney S (2015) How the University of Pittsburgh is using the NIST cybersecurity framework. https://www.sei.cmu.edu/podcasts/podcast_episode.cfm?episodeid=445056&autostarter=1&wtpodcast=howtheuniversityofpittsburghisusingthenistcybersecurityframework. Accessed 1 Feb 2018
32. University of Chicago (2016) Applying the cybersecurity framework at the university of Chicago: an education case study. http://security.bsd.uchicago.edu/wp-content/uploads/sites/2/2016/04/BSD-Framework-Implementation-Case-Study_final_edition.pdf. Accessed 31 Jan 2018

Course Code: CSL603	**Course Name: DWM LAB**
Class: TE-CO	**Batch: 3**
Roll no: 18CO63	**Name: SHAIKH TAUSEEF MUSHTAQUE ALI**

Assignment: 01

Aim: IEEE paper on selected topic

Data:

P.T.O

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 6, Special Issue , August 2019
International Conference on Recent Advances in Science, Engineering, Technology and Management at Sree Vahini Institute of Science and Technology-Tiruvuru, Krishna Dist, A.P

Data Mining Techniques For Big Data

Ahmed UnnisaBegum, Mohammed Ashfaq Hussain , Mubeena Shaik

Lecturer, Jazan University, Kingdom of Saudi Arabia
Research Scholar, Acharya Nagarjuna University, India

ABSTRACT: Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems to identify future patterns. Big Data is a term used for any data that is large in quantity. It is used to refer to any kind of data that is difficult to be represented using conventional methods like Database Management Systems or Microsoft Excel. In this paper we are discussing the characteristics applications of Big Data processing model and Big Data revolution, from the data mining view.

KEYWORDS: Data mining, Bid Data, Data Security, Data availability.

I. INTRODUCTION

Data mining refers to the activity of going through big data sets to look for relevant or appropriate information. The idea is that businesses collect massive sets of data that may be homogeneous or automatically collected[2]. Decision-makers need access to smaller, more specific pieces of data from those large sets. They use data mining to uncover the pieces of information that will inform leadership and help chart the course for a business. Data mining can involve the use of different kinds of software packages such as analytics tools. It can be automated, where individual workers send specific queries for information to database. Generally, data mining refers to operations that involve relatively sophisticated search operations that return targeted and specific results. For example, a data mining tool may look through dozens of years of accounting information to find a specific column of expenses or accounts receivable for a specific working year[1]. The importance of Big Data does not mean how much data we have but what would you get out of that data. We can analyze data to reduce cost and time, smart decision making.

II. DATA MINING AND BIG DATA

Data mining, also known as data discovery or knowledge discovery, is the process of analyzing data from different viewpoints and resulting it into useful information. This information is used by businesses to increase their revenue and reduce operational expenses. The software programs used in data mining are amongst the number of tools used in data analysis.The software enables users to analyze data from different point of views, classify it and make a summary of the data trends identified[3][4]. Technically, data mining involves the process of discovering patterns or relationships in large areas of related databases.The actual data mining task is the automatic or semi-automatic analysis of large datasets. This is done to assist in the extraction of previously unknown and unusual data patterns. These include detecting abnormalities in records, cluster analysis of data files and sequential pattern mining. Database techniques like spatial indices are commonly used in these processes.
After these processes, the patterns can be seen as the summary of the input data and can be used in further analysis like predictive analytics or machine learning. For instance, multiple groups of data can be identified through data mining steps. This is the process of analyzing larger data sets with the aim of uncovering useful information. Examples of this information include market trends, customer preferences, hidden patterns and unknown correlations. The analytics findings usually lead to new revenue opportunities, improved operational efficiency, more efficient marketing and other business benefits[2]. Companies often rely on big data analytics to assist them in making strategic business decisions. Big data analytics enable data scientists, predictive modelers and other professionals in the analytics field to analyze large volumes of transaction data. They can also use big data analytics to analyze data which might not have been discovered by conventional business programs.

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 6, Special Issue , August 2019
International Conference on Recent Advances in Science, Engineering, Technology and Management at Sree Vahini Institute of Science and Technology-Tiruvuru, Krishna Dist, A.P

III. CHALLENGES TO HANDLE BIG DATA

The programmers have to take decisions due to large availability of raw and complex data.An organization can collect, store, and analyze these large datasets in a number of ways. The Business can even use robust big data tools to store, access, and manage the structured and unstructured data collected from various sources in a faster and more efficient way. There are few challenges to address when handling big chunks of data. Some challenges listed below:

A. Handling a Large Amount of Data

The large availability of data makes the difficulty is making decisions.Data that enterprises can access has been increased exponentially from last several years. They have data for everything, right from what a consumer likes, to how they react, to a particular scent, to the amazing restaurant that opened up in Italy last weekend.
This data exceeds the amount of data that can be stored and computed, as well as retrieved. The challenge is not so much the availability, but the management of this data[5]. Along with rise in unstructured data, the availability of data is in multiple formats such as video, audio, social media, smart device data etc. Some of the newest ways developed to manage this data are a hybrid of relational databases combined with NoSQL databases. An example of this is MongoDB, which is an inherent part of the MEAN stack. There are also distributed computing systems like Hadoop to help manage Big Data volumes.

B. Data Security

In increasing of data, the major issue is to secure the data. Many organizations claim that they face trouble with Data Security. This happens to be a bigger challenge for them than many other data-related problems. The data that comes into enterprises is made available from a wide range of sources, some of which cannot be trusted to be secure and compliant within organizational standards.They need to use a variety of data collection strategies to keep up with data needs. This in turn leads to inconsistencies in the data, and then the outcomes of the analysis[6]. This data is made available from numerous sources, and therefore has potential security problems. You may never know which channel of data is compromised, thus compromising the security of the data available in the organization, and giving hackers a chance to move in. Now it is essential to introduce Data Security best practices for secure data collection, storage and retrieval.

1. Data Complexity

With the huge updating in data in every second, organizations need to be aware of handling it too. For example, if a retail company wants to analyze customer behaviour, real-time data from their current purchases can help. There are Data Analysis tools available for the same – Veracity and Velocity. They come with ETL engines, visualization, computation engines, frameworks and other necessary inputs.It is important for businesses to keep themselves updated with this data, along with the “stagnant” and always available data. This will help build better insights and enhance decision-making capabilities.

2. Shortage of Skilled Resources

There is a shortage of skilled Big Data professionals available at this time. This has become mentioned by many enterprises seeking to better utilize Big Data and build more effective Data Analysis systems. There is a lack experienced people and certified Data Scientists or Data Analysts available at present, which makes the “number crunching” difficult, and insight building slow.Again, training people at entry level can be expensive for a company dealing with new technologies. Many are instead working on automation solutions involving Machine Learning and Artificial Intelligence to build insights, but this also takes well-trained staff or the outsourcing of skilled developers.

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 6, Special Issue , August 2019
International Conference on Recent Advances in Science, Engineering, Technology and Management at Sree Vahini Institute of Science and Technology-Tiruvuru, Krishna Dist, A.P

IV. DATA MINING TECHNIQUES TO HANDLE BIG DATA V.

Data mining techniques have been around for many years in combination with data warehouses, and have now taken on greater prevalence with the advent of Big Data. Data analytics and the growth in both structured and unstructured data has also prompted data mining techniques to change, since companies are now dealing with larger data sets with more varied content. Additionally, artificial intelligence and machine learning are automating the process of data mining[7]. Some of the techniques which are listed below:

A. Association

Association makes a correlation between two or more items to identify a pattern. For instance, a supermarket could determine that customers often purchase whipped cream when they buy strawberries and vice versa. Association is often used at point-of-sale systems to determine common tendencies among products. "It's a very simple method, but you'd be surprised how much intelligence and insight it can provide—the kind of information many businesses use on a daily basis to improve efficiency and generate revenue," according to technology company Galvanize. Application areas include physical organization of items, marketing and the cross-selling and up-selling of products.

B. Classification: Multiple attributes can be used to identify a particular class of items. Classification assigns items into target categories or classes to accurately predict what will occur within the class. Several industries use classification with customers. For instance, a banking company could use a classification model to identify loan applicants as low, medium or high credit risks. Other organizations classify current and target audiences into age and social groups for marketing campaigns.

C. Clustering: "Clustering is the method by which like records are grouped together," according to Alex Berson, Stephen Smith and Kurt Thearling in the book Building Data Mining Applications for CRM. "Usually this is done to give the end user a high level view of what is going on in the database." Seeing object groupings can help businesses in areas like marketing segmentation. Clustering can be used in this example to subdivide a market into subsets of customers[7][8]. Each subset can then be targeted with a specific marketing strategy based on the attributes of the cluster, such as buying patterns for customers in one cluster vs. another cluster.

D. Decision Trees: Decision trees are used to categorize or predict data. A decision tree starts with a simple question that has two or more answers. Each answer leads to a further question that is used to classify or identify data that can be categorized, or so that a prediction can be made based on each answer.The graphic of a decision tree represents how a cell phone provider might classify customers who combine, or those who don't renew their phone contracts. The authors of Building Data Mining Applications for CRM offer some interesting takeaways for the graphic. It divides the data into each branch without losing any of the data. For instance, the total number of records in a parent node is equal to the sum of the records contained in its two children.

E. Sequential Patterns: Sequential patterns identify trends or regular occurrences of similar events. This data mining technique is often used to understand user buying behaviors. Many retailers use data and sequential patterns to decide on the products they display[6]. "With customer data you can identify that customers buy a particular collection of products together at different times of the year," according to IBM. "In a shopping basket application, you can use this information to automatically suggest that certain items be added to a basket based on their frequency and past purchasing history."

VI. CONCLUSION

In this paper we discussed the basic data mining techniques to handle Big data. Big data is evolving with the exponential rise in data availability. It is important for the organisations to work around these challenges and gain advantages over their competition with more reliable insights,. Thus, more studies should be addressed towards big data analytics to manage business with huge amount of data.

REFERENCES:

[1] S. San M. Negnevitsky, N. Hatziargyriou, "Applications of Data Mining and Analysis Techniques in Wind Power Systems", 42440178X/06/$20.00 ©2006 IEEE.
[2] Krioukov, Andrew, "Integrating Renewable Energy Using Data Analytics Systems: Challenges and Opportunities." IEEE Data Eng. Bull. 34.1 (2011): 3-11.

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 6, Special Issue , August 2019

International Conference on Recent Advances in Science, Engineering, Technology and Management at Sree Vahini Institute of Science and Technology-Tiruvuru, Krishna Dist, A.P

[3] Fayyad, U. Data Mining and Knowledge Discovery: Making Sense Out of IEEE Expert, v. 11, no. 5, pp. 20-25, October 1996.
[4] Chen, H., Chaing, R.H.L. and Storey, V.C. (2012) Business Intelligence and Analytics: From Big Data to Big Impact, MIS Quarterly, 36, 4, pp. 1165-1188.
[5] Tekiner F. and Keane J.A., Systems, Man and Cybernetics (SMC), "Big Data Framework" 2013 IEEE International Conference on 13-16 Oct. 2013, 1494-1499.
[6] Hsinchun Chen, Roger H. L. Chiang, Veda C. Storey, "Business Intelligence And Analytics: From Big Data To Big Impact, Big Data Analytics An Oracle White Paper", MIS Quarterly vol. 36 no. 4, pp. 1165-1188/December 2012.
[7] Anastasia, February 2015. Big data and new product development. Entrepreneurial Insights http://www.entrepreneurial-insights.com/big-data-new-product-development/, Accessed on June 15, 2015.
[8] Sagiroglu, S.; Sinanc, D.,"Big Data: A Review",2013,20-24.

Course Code: CSL603	**Course Name: DWM LAB**
Class: TE-CO	**Batch: 3**
Roll no: 18CO63	**Name: SHAIKH TAUSEEF MUSHTAQUE ALI**

Assignment: 02

Aim: Case Study on the given topic

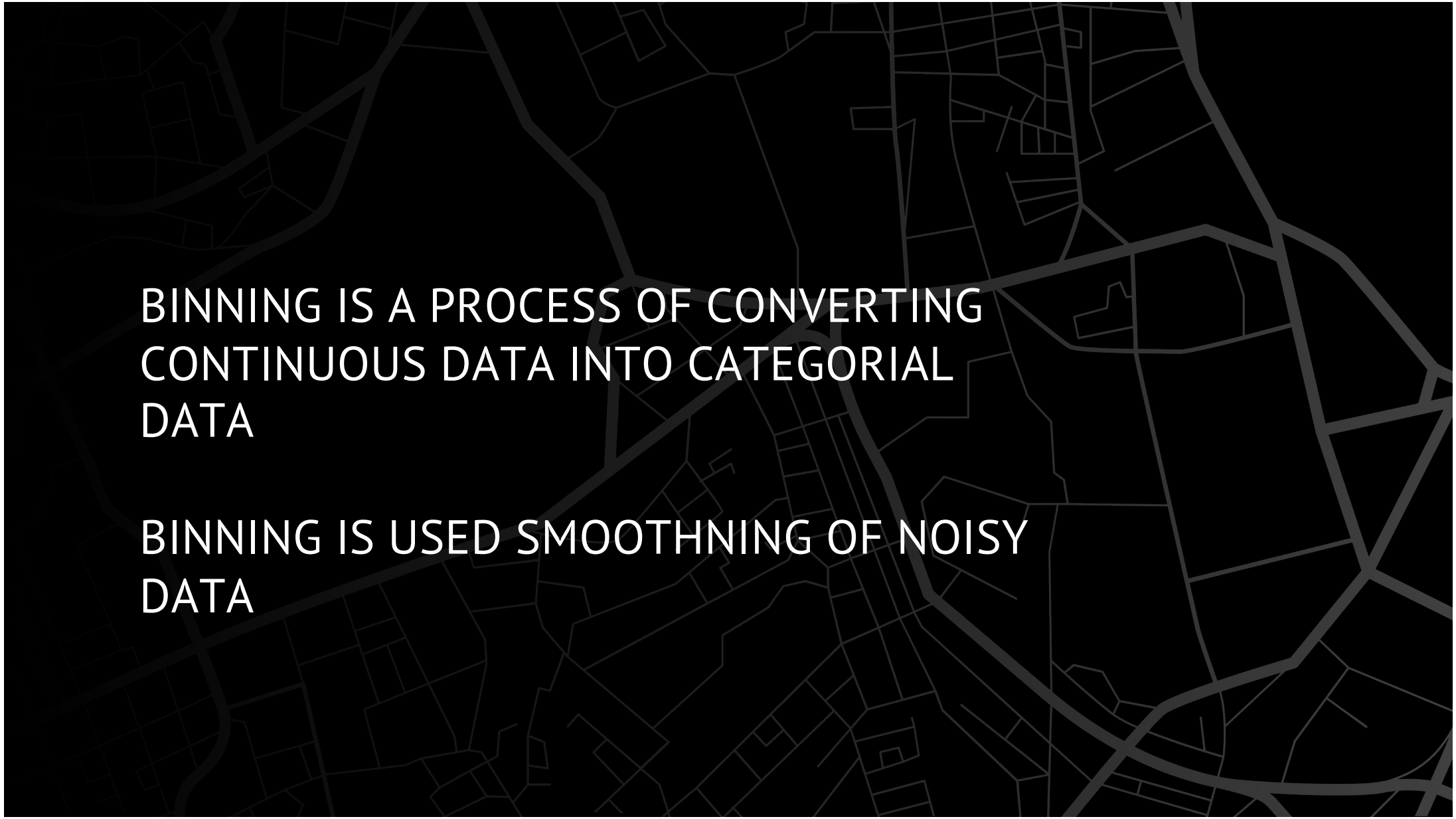
Data:

P.T.O


BINNING
Data Transformation & Data Discretization
From, 18CO63, 19DCO06 [GROUP NO. 15]

TABLE OF CONTENTS

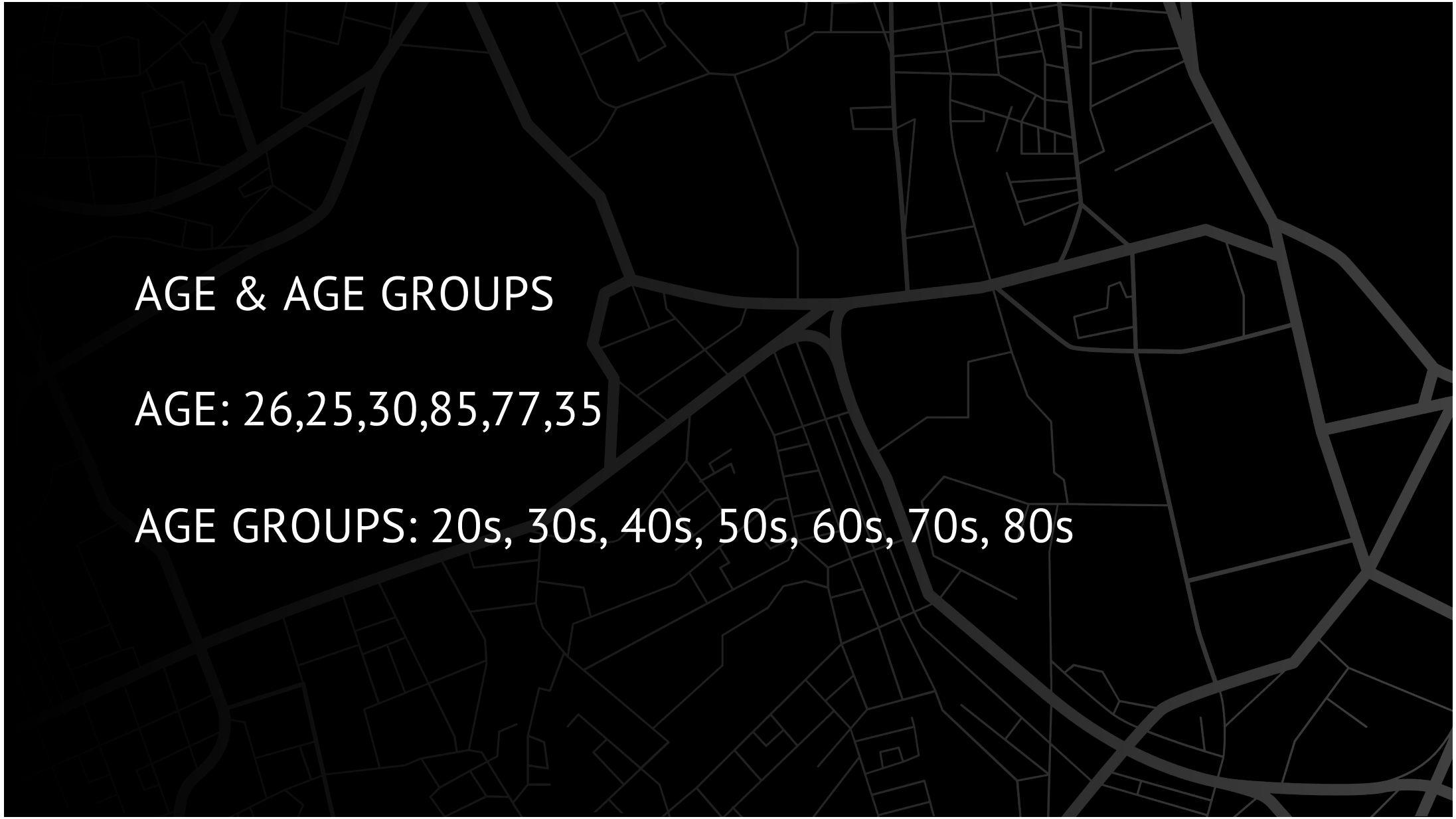

WHAT IS BINNING
01


BINNING IS A PROCESS OF CONVERTING CONTINUOUS DATA INTO CATEGORIAL DATA
BINNING IS USED SMOOTHNING OF NOISY DATA


WHAT IS BINS
02

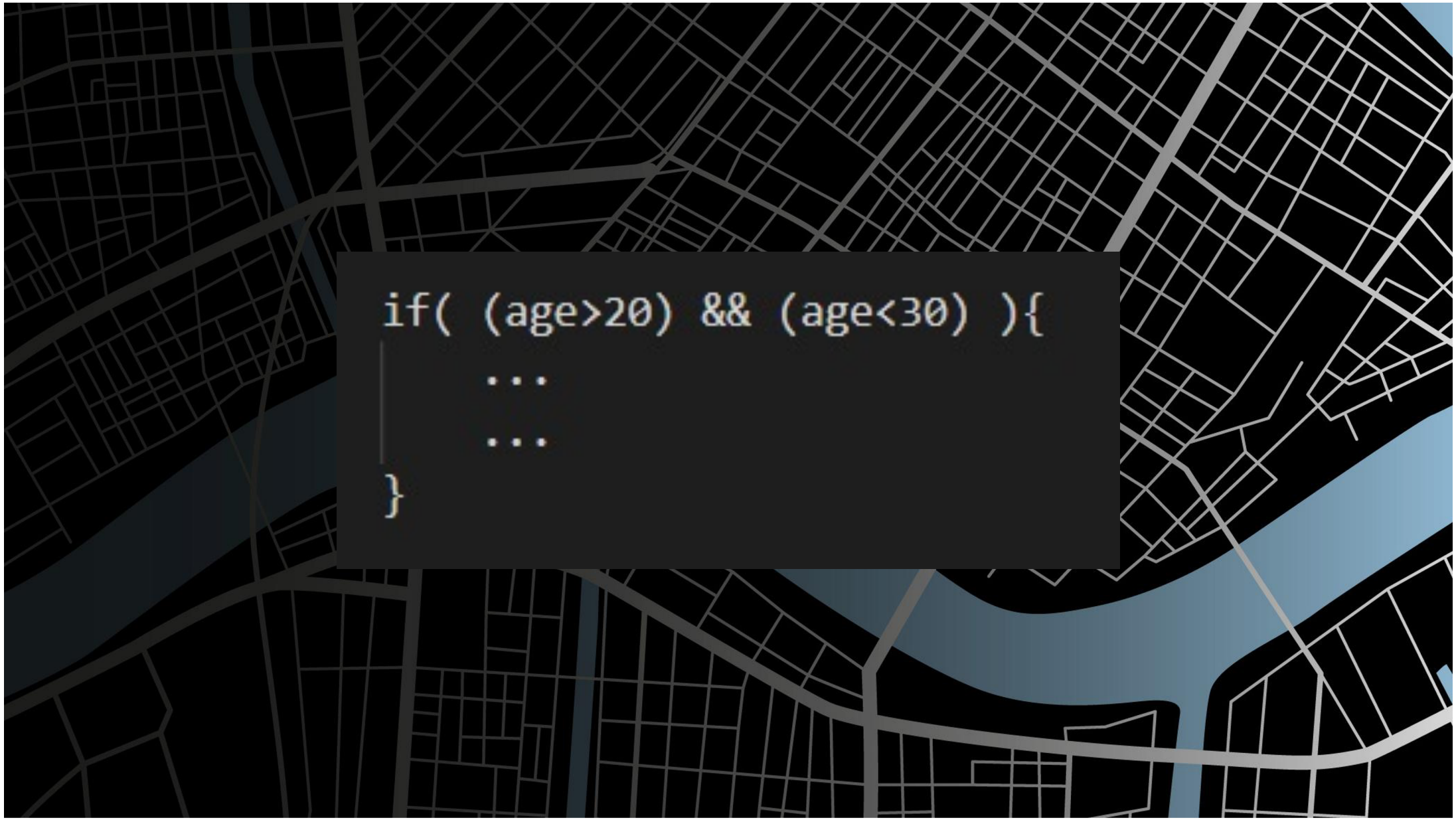
BINS ARE LOGICAL GROUPING OF DATA ACCORDING TO OUR PREDICTION AND PROVIDED DATA

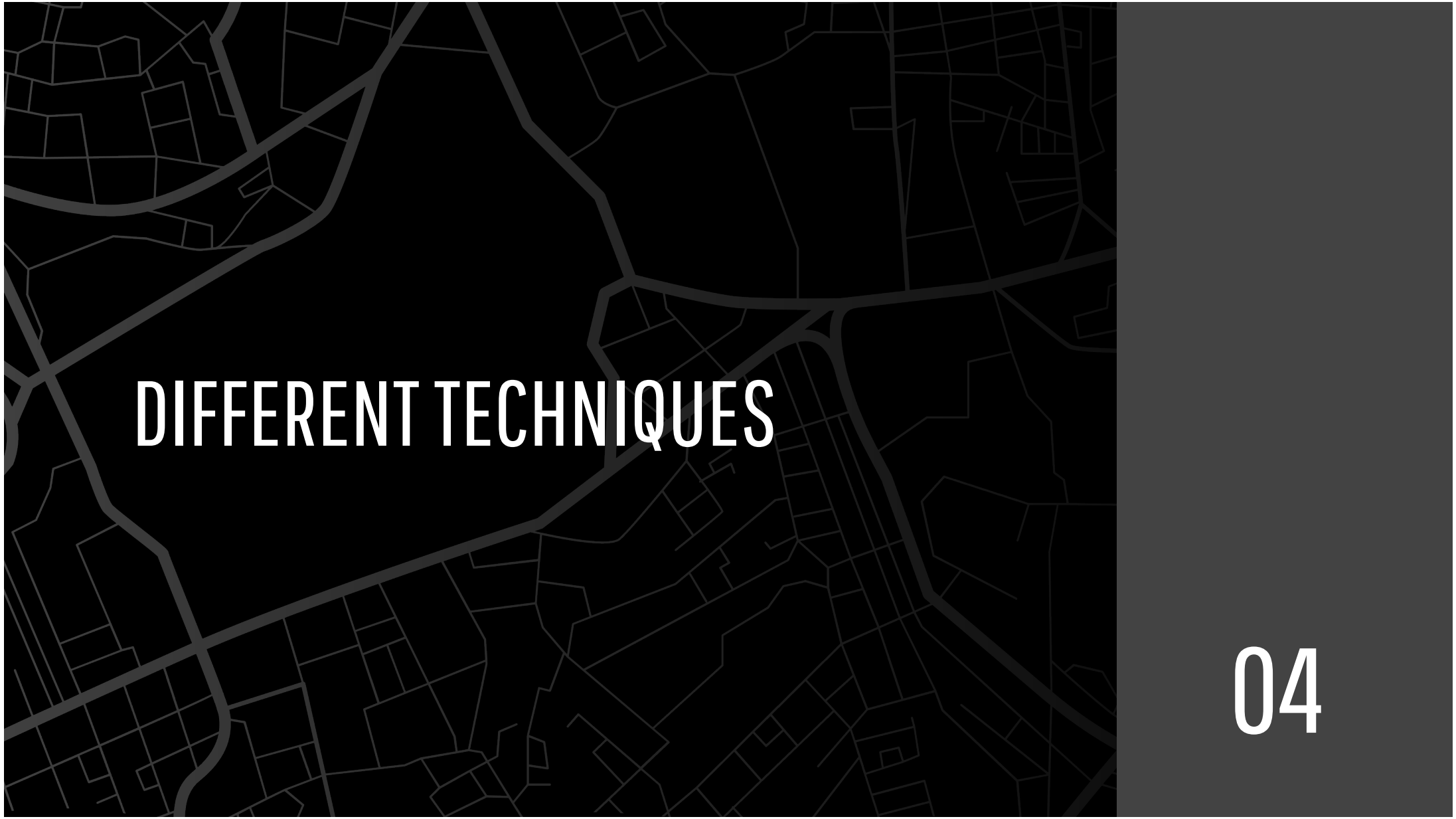

EXAMPLE
03


AGE & AGE GROUPS
AGE: 26,25,30,85,77,35
AGE GROUPS: 20s, 30s, 40s, 50s, 60s, 70s, 80s

PREDICTED DATA CATEGORIZATION

AGE	AGE GROUPS
26	20s
25	20s
30	30s
85	80s
77	70s
35	30s


if((age>20) && (age<30)){
...
...
}


DIFFERENT TECHNIQUES
04

DIFFERENT TECHNIQUES TO FORM A BIN

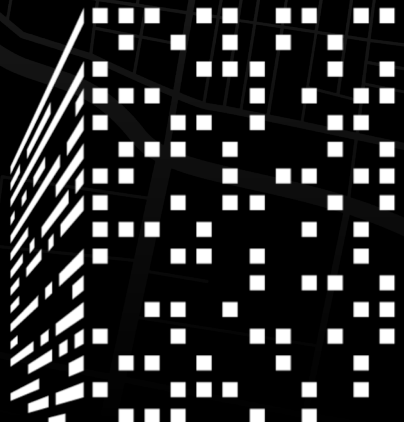
DATA: 4, 8, 15, 21, 21, 24, 25, 28, 34

EQUAL PARTITIONED BIN

BIN 1: 4, 8, 15
BIN 2: 21, 21, 24
BIN 3: 24, 28, 34

BIN MEAN

BIN 1: 9, 9, 9
BIN 2: 22, 22, 22
BIN 3: 29, 29, 29



BIN BOUNDRIES

BIN 1: 4, 4, 15
BIN 2: 21, 21, 24
BIN 3: 24, 25, 34

THANK YOU!

From, 18CO63, 19DCO06 [GROUP NO. 15]