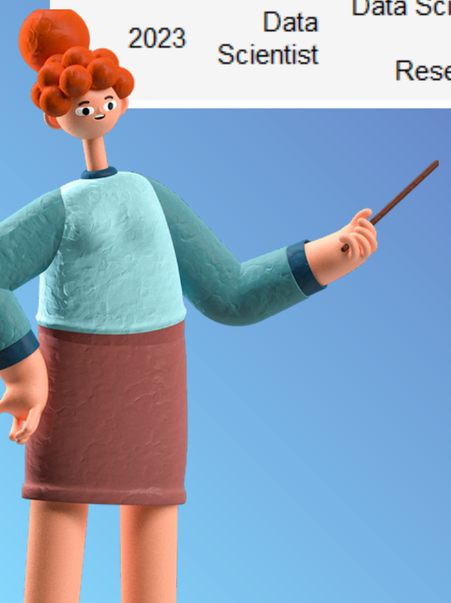
Jobs and Salaries In Data Science

Presentation by Charinrat , Thanwadee

ตารางข้อมูลทั้งหมด



work_year	job_title	job_category	salary_currency	salary	salary_in_usd	employee_residence	experience_level	employment_type	work_setting	company_location	company_size
2023	Data DevOps Engineer	Data Engineering	EUR	88000	95012	Germany	Mid-level	Full-time	Hybrid	Germany	L
2023	Data Architect	Data Architecture and Modeling	USD	186000	186000	United States	Senior	Full-time	In-person	United States	M
2023	Data Architect	Data Architecture and Modeling	USD	81800	81800	United States	Senior	Full-time	In-person	United States	M
2023	Data Scientist	Data Science and Research	USD	212000	212000	United States	Senior	Full-time	In-person	United States	M
2023	Data Scientist	Data Science and Research	USD	93300	93300	United States	Senior	Full-time	In-person	United States	M

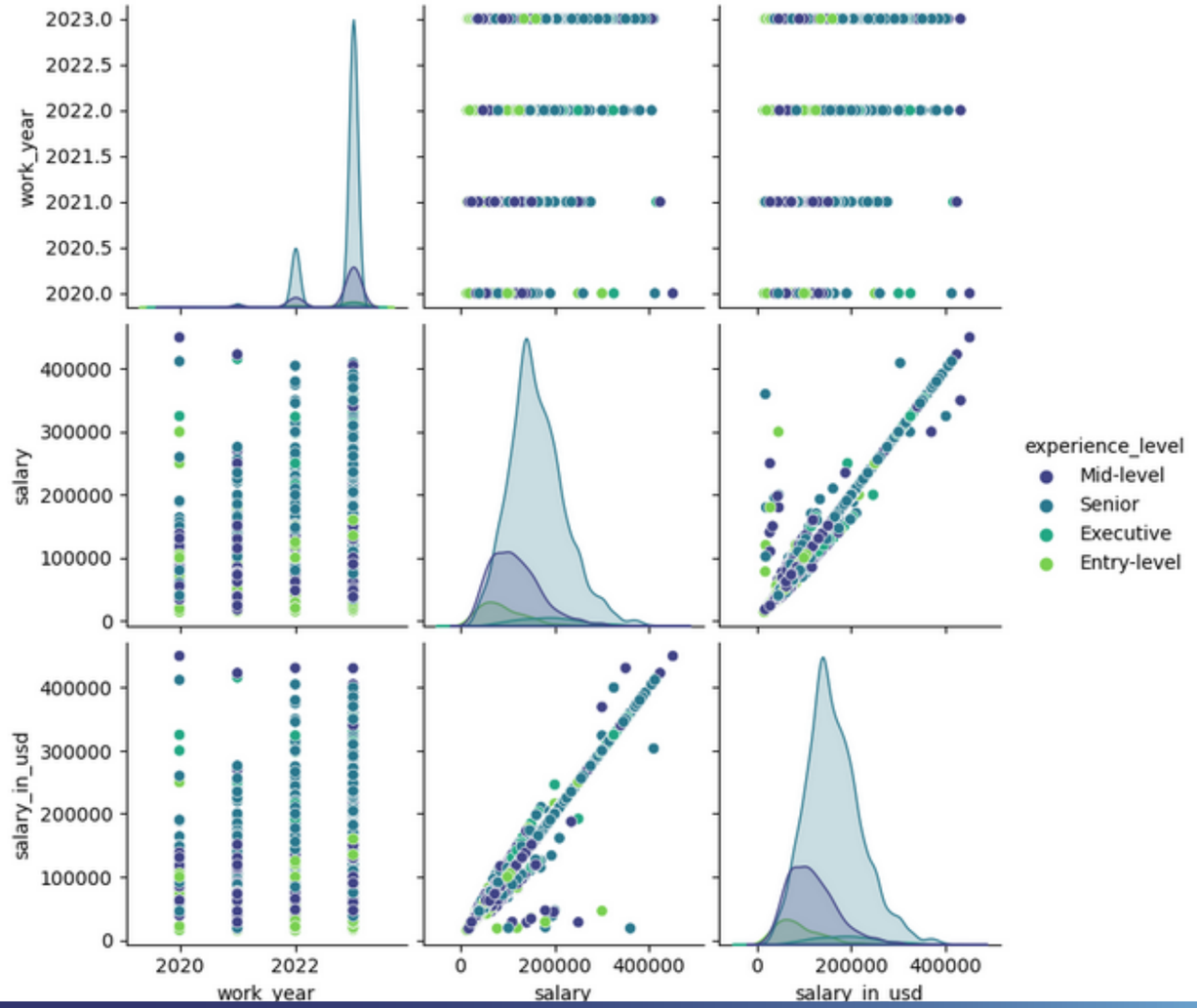
PAIRPLOT

เป็นการ **pLot** ข้อมูลทั้งหมดในตารางให้อยู่ในรูปแบบกราฟต่าง ๆ และแบ่งประเภทตาม **experience_level**

```
]: sns.pairplot(J,hue='experience_level' ,palette="viridis")
```

```
C:\Users\User\anaconda3\Lib\site-packages\seaborn\axisgrid.py:118: UserWarning: The figure layout has changed to tight
  self._figure.tight_layout(*args, **kwargs)
```

```
Out[16]: <seaborn.axisgrid.PairGrid at 0x2a4f9b4b650>
```

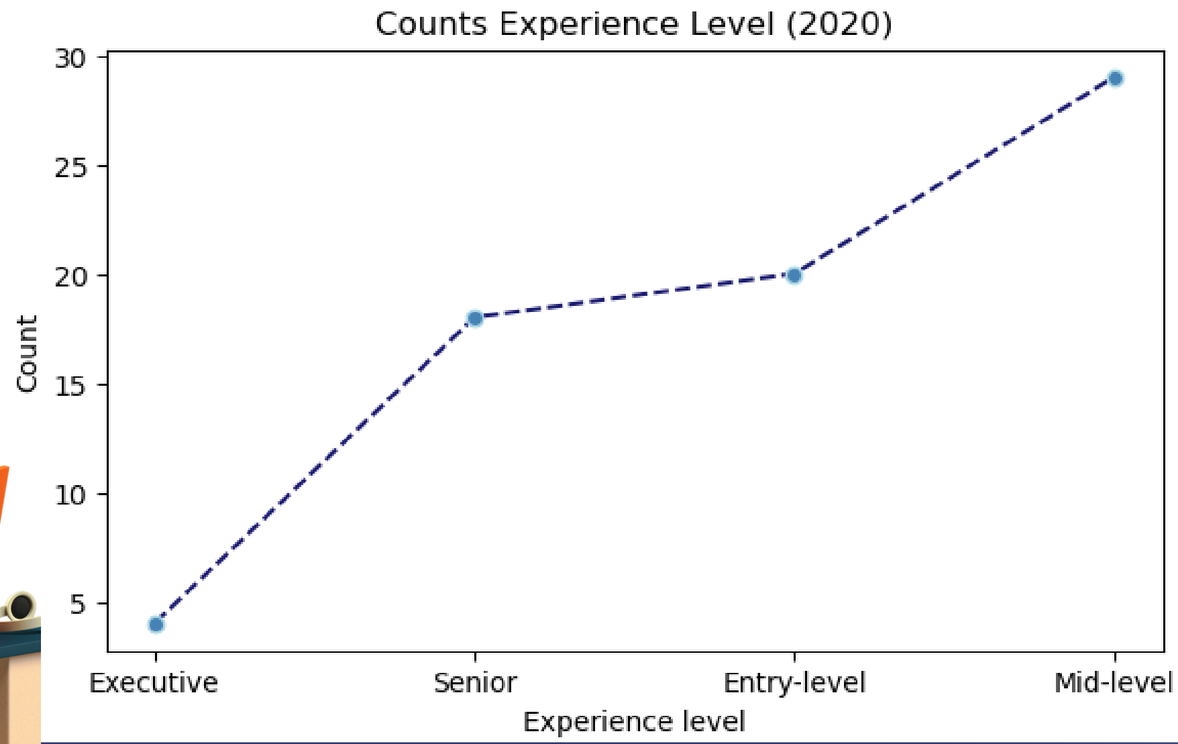


PLOT

```
In [38]: JW = J[J['work_year']==2020]['experience_level'].value_counts()
JW.sort_values(ascending=True, inplace=True)
x = JW.index
y = JW

plt.figure(figsize=[7,4])
plt.plot(x,y,'--',marker='o',mec='#ADD8E6',mfc='#4682B4',color='#191970')
plt.title('Counts Experience Level (2020)')
plt.xlabel('Experience level')
plt.ylabel('Count')

plt.show()
```



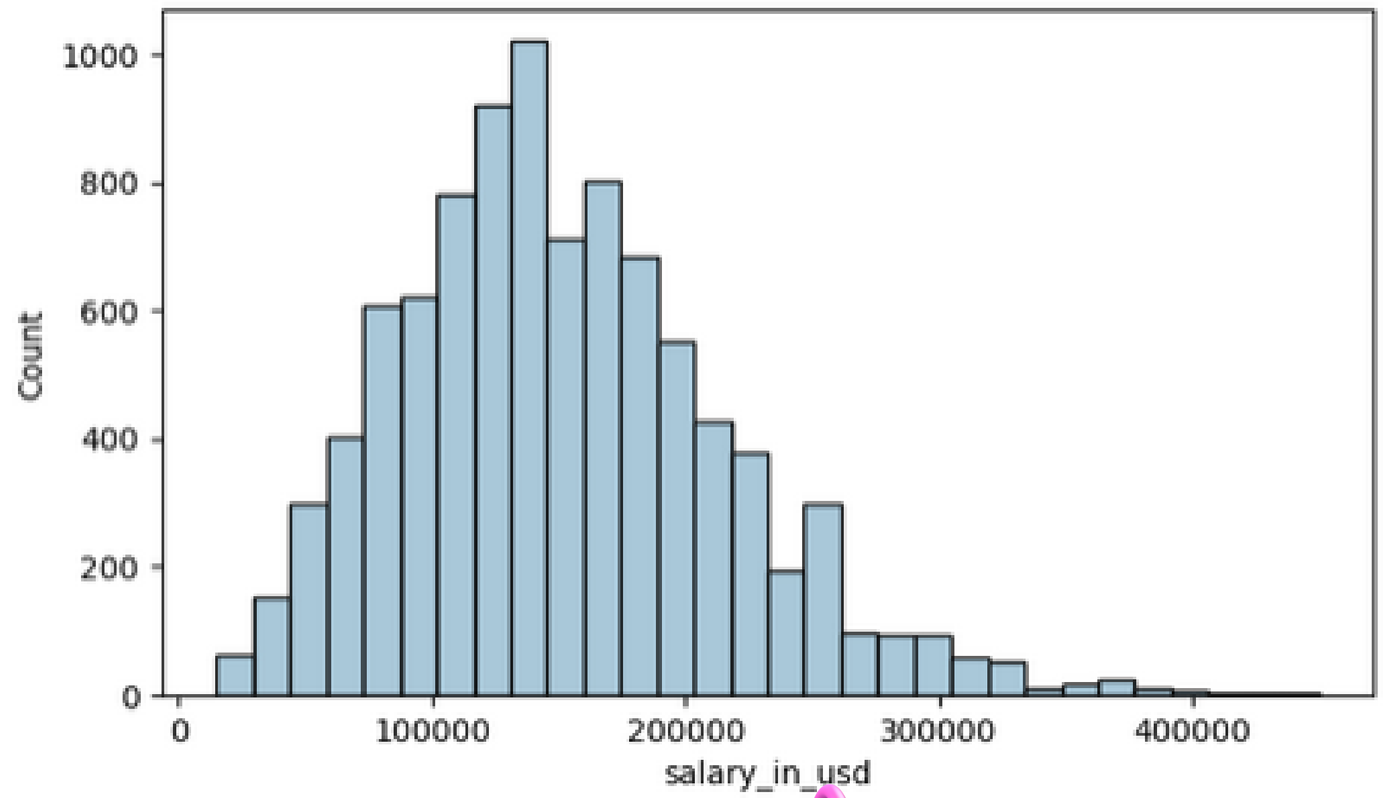
จะใช้การ **plot** ที่มี**column work_ year & experience_ level** โดยแสดงข้อมูลปี **2020**

HISTOGRAM

จะใช้การ **plot** ด้วย **seaborn** แบบ **hist** ที่มี **column salary_in_usd** ในการแสดงช่วงเงินที่ได้รับมากที่สุด

```
In [11]: plt.figure(figsize=[7,4])
         sns.histplot(J['salary_in_usd'],bins=30,color ="#8DB6CD")
```

Out[11]: <Axes: xlabel='salary_in_usd', ylabel='Count'>



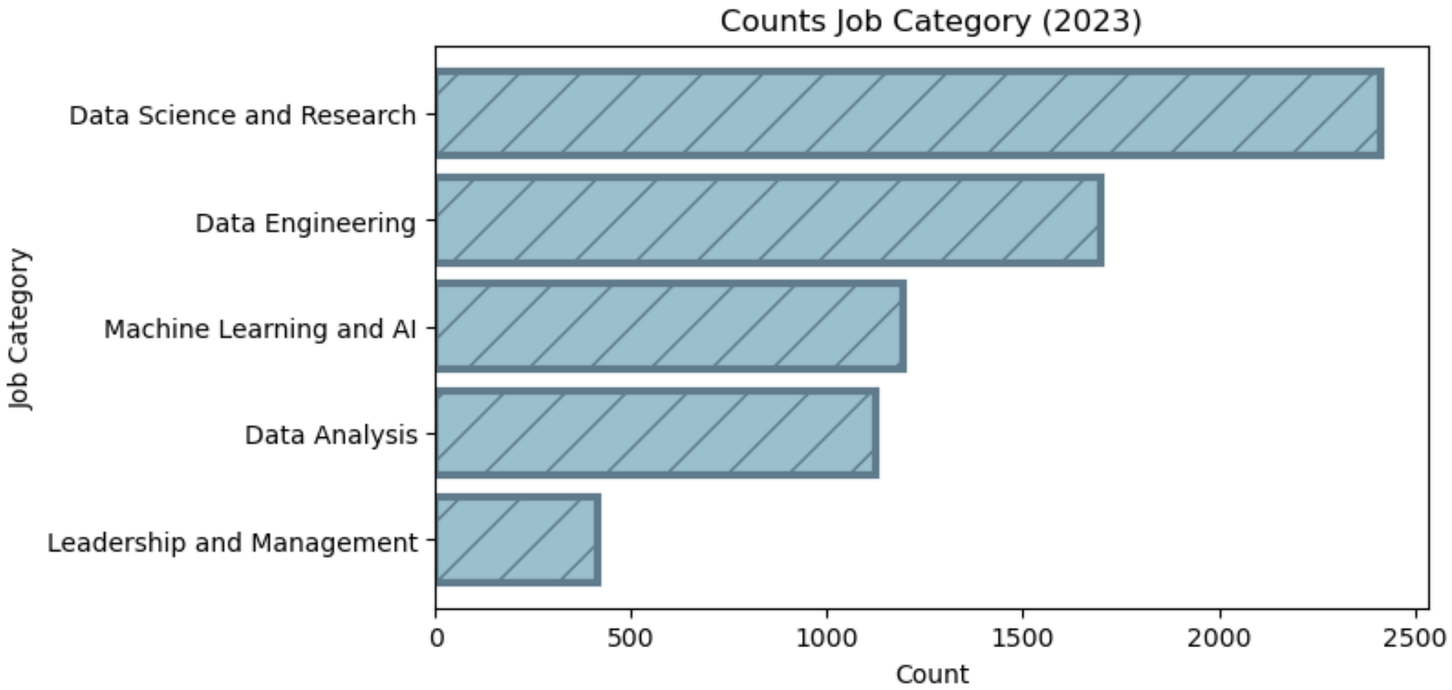
BAR CHARTS

เป็นการ **plot** แบบ **bar** ที่แสดงข้อมูลของ **job_category** ที่มีคนทำมากที่สุดของปี **2023** จาก **5** อันดับแรก

```
In [37]: J1 = J[J['work_year']==2023]['job_category'].value_counts().head().sort_values()
         x = J1.index
         y = J1

         plt.figure(figsize=[7,4])
         plt.barh(x,y,color='#9AC0CD',hatch='/',ec='#607B8B',lw=3)
         plt.title('Counts Job Category (2023)')
         plt.xlabel('Count')
         plt.ylabel('Job Category')

         plt.show()
```



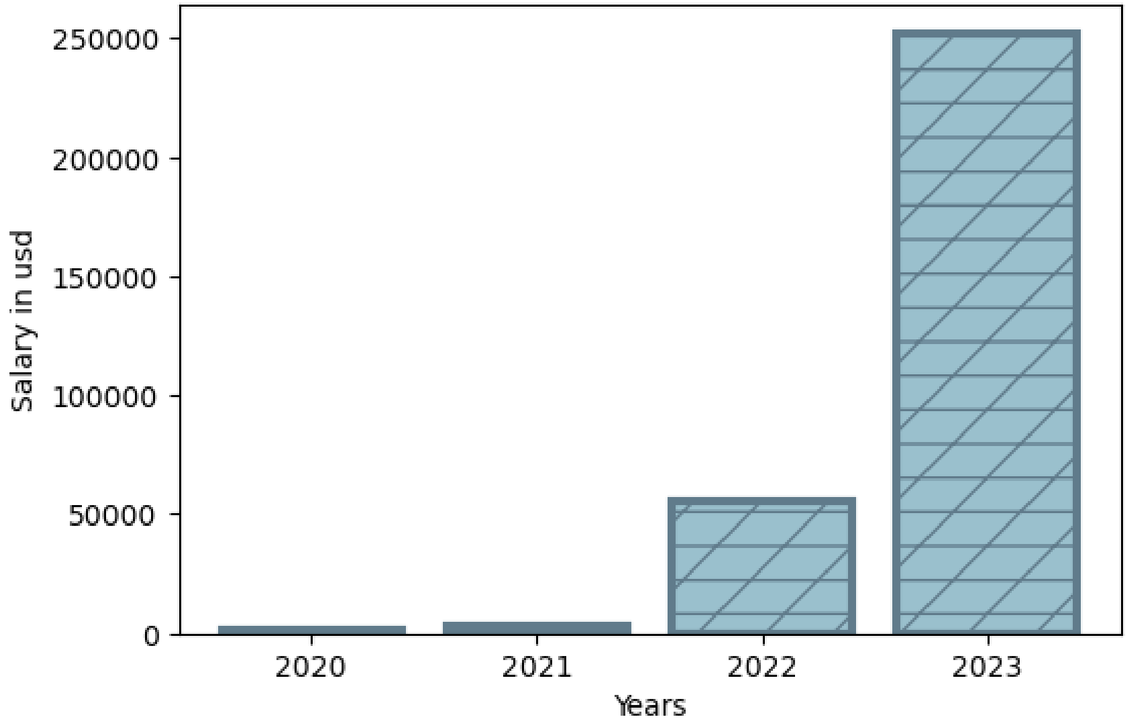
BAR CHARTS

เป็นการ **plot** แบบ **bar** ที่แสดงข้อมูลเงินเดือนรวมของอาชีพ **Data Scientist** ในทั้ง **4** ปี

```
In [39]: x = arr_J['Data Scientist'].index
         y = arr_J['Data Scientist']//1000

         plt.figure(figsize=[6,4])
         plt.bar(x,y,color='#9AC0CD',hatch='/-',ec='#607B8B',lw=3)
         plt.xticks([2020,2021,2022,2023])
         plt.xlabel('Years')
         plt.ylabel('Salary in usd')

         plt.show()
```



CLASSIFICATION

In [12]:
```
att = J[['salary_in_usd', 'work_year']]
label = J['experience_level']

att_train, att_test, label_train, label_test = train_test_split(att, label, random_state = 0, train_size = 0.75)

scaler = StandardScaler()
scaler.fit(att_train)
att_train[[ 'salary_in_usd', 'work_year']] = scaler.transform(att_train)

model = KNeighborsClassifier(n_neighbors = 3)
model.fit(att_train, label_train)

print(model.score(scaler.transform(att_test), label_test))

result = pd.concat([att_test,label_test],axis =1)
result['predict'] = model.predict(scaler.transform(att_test))
result
```

```
C:\Users\User\anaconda3\Lib\site-packages\sklearn\base.py:464: UserWarning
ted with feature names
  warnings.warn(
```

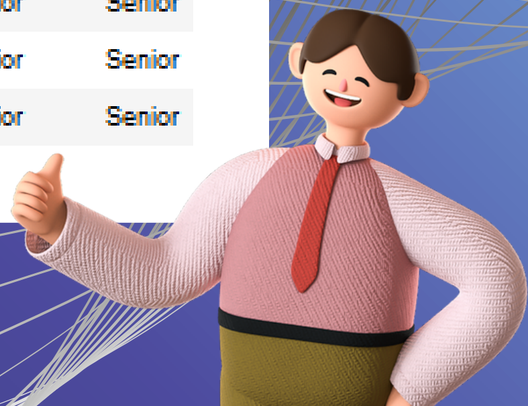
0.6840530141085934

Out[12]:

	salary_in_usd	work_year	experience_level	predict
8584	95000	2022	Senior	Senior
8027	225000	2022	Senior	Senior
4194	18381	2023	Senior	Entry-level
5899	200000	2023	Senior	Senior
587	110000	2023	Senior	Mid-level
...	...	...	...	...
8670	188100	2022	Senior	Senior
1413	122400	2023	Senior	Senior
8825	164996	2022	Senior	Senior
4160	166364	2023	Senior	Senior
5667	145000	2023	Senior	Senior

2339 rows × 4 columns

ใช้ **Classification** ทำนายหา **experience_level** จาก **column salary_in_usd** และ **work_year** ที่ความแม่นยำประมาณ **0.6841**



CLASSIFICATION


```
In [54]: att = J[['salary_in_usd', 'work_year']]
label = J['company_size']

att_train, att_test, label_train, label_test = train_test_split(att, label, random_state = 0, train_size = 0.75)

scaler = StandardScaler()
scaler.fit(att_train)
att_train[['salary_in_usd', 'work_year']] = scaler.transform(att_train)

model = KNeighborsClassifier(n_neighbors = 3)
model.fit(att_train, label_train)

print(model.score(scaler.transform(att_test), label_test))

result = pd.concat([att_test,label_test],axis =1)
result['predict'] = model.predict(scaler.transform(att_test))
result
```

```
C:\Users\User\anaconda3\Lib\site-packages\sklearn\base.py:464: UserWarning: X does
tted with feature names
  warnings.warn(
C:\Users\User\anaconda3\Lib\site-packages\sklearn\base.py:464: UserWarning: X does
tted with feature names
  warnings.warn(
```

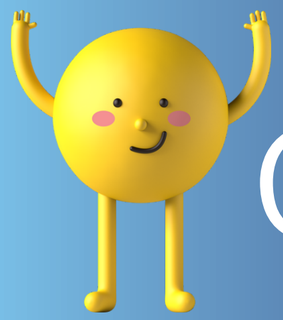
0.9144933732364259

Out[54]:

	salary_in_usd	work_year	company_size	predict
8584	95000	2022	M	M
8027	225000	2022	M	M
4194	18381	2023	L	L
5899	200000	2023	M	M
587	110000	2023	M	M
...	...	...	...	...
8670	188100	2022	M	M
1413	122400	2023	M	M
8825	164996	2022	M	M
4160	166364	2023	M	M
5667	145000	2023	M	M

2339 rows × 4 columns

ใช้ **Classification** ทำนายหา **company_size** จาก **column salary_in_usd** และ **work_year** ที่ความแม่นยำประมาณ **0.9145**



Group members

- น.ส.ชรินทร์รัตน์ เมืองจันทร์ 6321651591
- น.ส.ธัญวดี เคลือบมงคล 6421650392