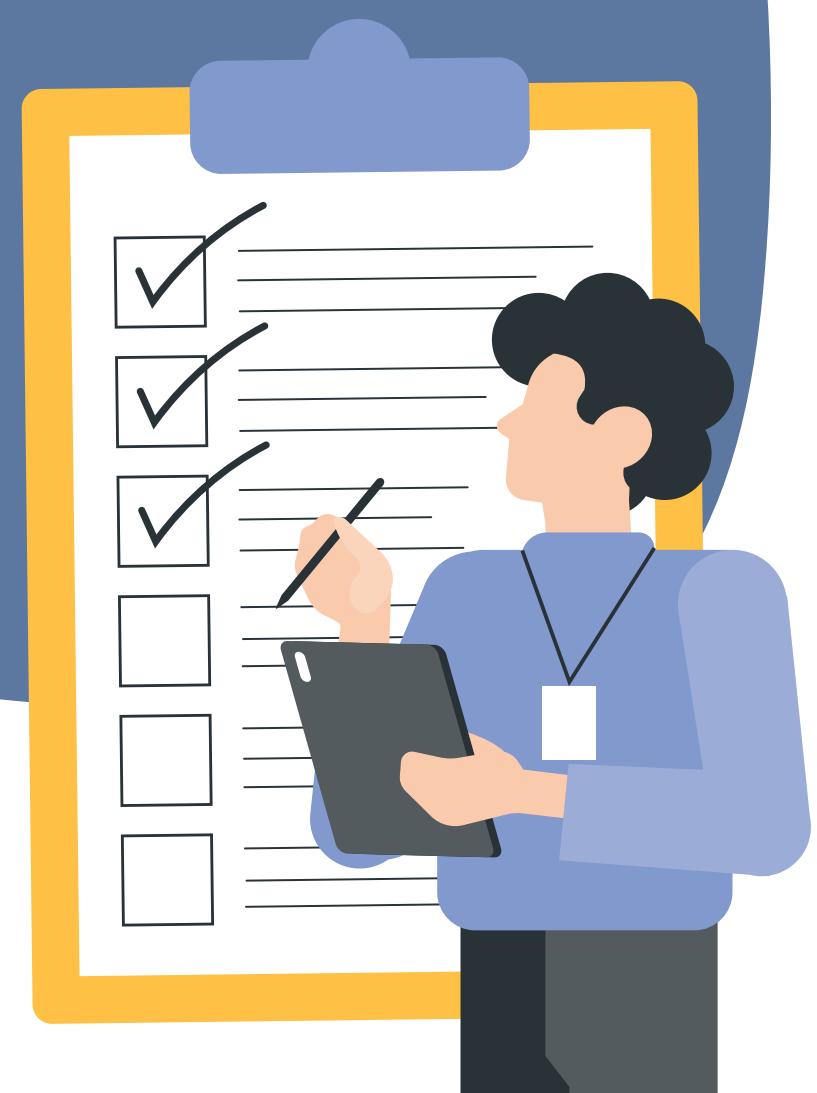


# Exploratory Data Analysis (EDA) on Lead Scoring Dataset

GROUP 20

EDA on Lead Scoring Dataset



# EDA คืออะไร?

คือกระบวนการสำรวจข้อมูลเบื้องต้นด้วยวิธีการต่าง ๆ  
เพื่อกำความเข้าใจข้อมูลก่อนจะดำเนินการวิเคราะห์หรือสร้างโมเดลทางสถิติ



# EDA มีความสำคัญกับ Data Mining อย่างไร?

---

การทำ EDA ช่วยให้กระบวนการวิเคราะห์ข้อมูลมีความแม่นยำและประสิทธิภาพสูงขึ้น



- นำเข้าและการเตรียมการใช้ข้อมูล
- รายละเอียดโครงสร้างและขนาดของชุดข้อมูล
- จัดการและทำความสะอาดข้อมูล
- ปรับแต่งการแสดงผลและแสดงภาพข้อมูล
- วิเคราะห์ข้อมูลเชิงลึก EDA

# ข้อมูลที่นำมาศึกษา

ข้อมูลนี้ได้มาจากการบริษัทขายหลักสูตรออนไลน์เกี่ยวกับอุตสาหกรรม  
ที่ได้จัดเก็บข้อมูลการจัดการลูกค้าเป้าหมาย (Leads)





EDA

# การนำเข้าข้อมูลและเตรียมการใช้ข้อมูล

LOADING...



# Import ข้อมูลจาก kaggle เข้า Google Colab

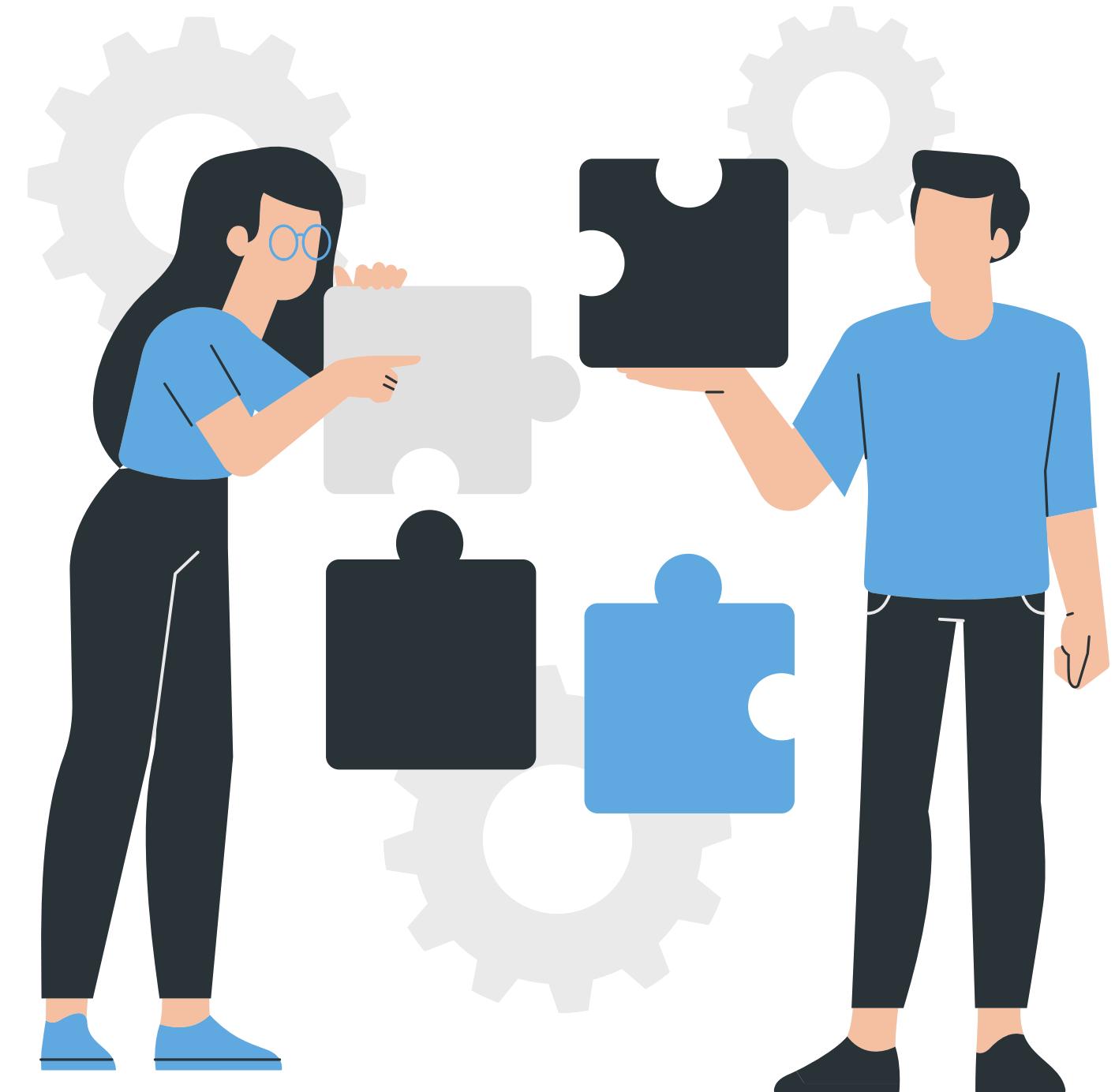
```
[ ] !pip install kaggle

from google.colab import files

import os
os.environ['KAGGLE_CONFIG_DIR'] = '/content'

!kaggle datasets download -d ashhydv/leads-dataset

import zipfile
with zipfile.ZipFile('leads-dataset.zip', 'r') as zip_ref:
    zip_ref.extractall('/content/leads-dataset')
```



# Import library

```
[ ] #Importing Libraries  
import numpy as np  
import pandas as pd  
import matplotlib.pyplot as plt  
import seaborn as sns
```

---

---

---

## ไม่ให้โปรแกรมแสดง warning

```
[ ] #Suppress Warnings  
import warnings  
warnings.filterwarnings("ignore")
```

- numpy : ใช้สำหรับการคำนวณเชิงคณิตศาสตร์
- pandas : ใช้สำหรับการจัดการและการวิเคราะห์ข้อมูลในรูปแบบของ DataFrame
- matplotlib.pyplot : ใช้สำหรับการสร้างกราฟและการแสดงผลข้อมูล
- seaborn : ใช้สำหรับการสร้างกราฟที่มีความซับซ้อนและสวยงามมากขึ้น

EDA

# รายละเอียดของโครงสร้าง และขนาดของชุดข้อมูล



# Loading Dataset

```
▶ #Loading Dataset  
data = pd.read_csv('/content/leads-dataset/Leads.csv')
```

LOADING...

## GROUP 20

```
#Loading Dataset
data = pd.read_csv('/content/leads-dataset/Leads.csv')

data.info()
```

<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 9240 entries, 0 to 9239  
Data columns (total 37 columns):  
 # Column Non-Null Count Dtype  
--- ---  
 0 Prospect ID 9240 non-null object  
 1 Lead Number 9240 non-null int64  
 2 Lead Origin 9240 non-null object  
 3 Lead Source 9204 non-null object  
 4 Do Not Email 9240 non-null object  
 5 Do Not Call 9240 non-null object  
 6 Converted 9240 non-null int64  
 7 TotalVisits 9103 non-null float64  
 8 Total Time Spent on Website 9240 non-null int64  
 9 Page Views Per Visit 9103 non-null float64  
 10 Last Activity 9137 non-null object  
 11 Country 6779 non-null object  
 12 Specialization 7802 non-null object  
 13 How did you hear about X Education 7033 non-null object  
 14 What is your current occupation 6550 non-null object  
 15 What matters most to you in choosing a course 6531 non-null object  
 16 Search 9240 non-null object  
 17 Magazine 9240 non-null object  
 18 Newspaper Article 9240 non-null object  
 19 X Education Forums 9240 non-null object  
 20 Newspaper 9240 non-null object  
 21 Digital Advertisement 9240 non-null object  
 22 Through Recommendations 9240 non-null object  
 23 Receive More Updates About Our Courses 9240 non-null object  
 24 Tags 5887 non-null object  
 25 Lead Quality 4473 non-null object  
 26 Update me on Supply Chain Content 9240 non-null object  
 27 Get updates on DM Content 9240 non-null object  
 28 Lead Profile 6531 non-null object  
 29 City 7820 non-null object  
 30 Asymmetrique Activity Index 5022 non-null object  
 31 Asymmetrique Profile Index 5022 non-null object  
 32 Asymmetrique Activity Score 5022 non-null float64  
 33 Asymmetrique Profile Score 5022 non-null float64  
 34 I agree to pay the amount through cheque 9240 non-null object  
 35 A free copy of Mastering The Interview 9240 non-null object  
 36 Last Notable Activity 9240 non-null object  
dtypes: float64(4), int64(3), object(30)  
memory usage: 2.6+ MB

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9240 entries, 0 to 9239
Data columns (total 37 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   Prospect ID     9240 non-null   object 
 1   Lead Number     9240 non-null   int64  
 2   Lead Origin     9240 non-null   object 
 3   Lead Source     9204 non-null   object 
 4   Do Not Email    9240 non-null   object 
 5   Do Not Call     9240 non-null   object 
 6   Converted       9240 non-null   int64  
 7   TotalVisits     9103 non-null   float64
 8   Total Time Spent on Website 9240 non-null   int64  
 9   Page Views Per Visit 9103 non-null   float64
 10  Last Activity   9137 non-null   object 
 11  Country         6779 non-null   object 
 12  Specialization  7802 non-null   object 
 13  How did you hear about X Education 7033 non-null   object 
 14  What is your current occupation 6550 non-null   object 
 15  What matters most to you in choosing a course 6531 non-null   object 
 16  Search          9240 non-null   object 
 17  Magazine        9240 non-null   object 
 18  Newspaper Article 9240 non-null   object 
 19  X Education Forums 9240 non-null   object 
 20  Newspaper       9240 non-null   object 
 21  Digital Advertisement 9240 non-null   object 
 22  Through Recommendations 9240 non-null   object 
 23  Receive More Updates About Our Courses 9240 non-null   object 
 24  Tags            5887 non-null   object 
 25  Lead Quality    4473 non-null   object 
 26  Update me on Supply Chain Content 9240 non-null   object 
 27  Get updates on DM Content 9240 non-null   object 
 28  Lead Profile    6531 non-null   object 
 29  City            7820 non-null   object 
 30  Asymmetrique Activity Index 5022 non-null   object 
 31  Asymmetrique Profile Index 5022 non-null   object 
 32  Asymmetrique Activity Score 5022 non-null   float64
 33  Asymmetrique Profile Score 5022 non-null   float64
 34  I agree to pay the amount through cheque 9240 non-null   object 
 35  A free copy of Mastering The Interview 9240 non-null   object 
 36  Last Notable Activity 9240 non-null   object 
dtypes: float64(4), int64(3), object(30)
memory usage: 2.6+ MB
```

EDA  
การจัดการ  
และ  
ทำความสะอาดข้อมูล



# Dropping Columns

ลบข้อมูลที่ไม่สำคัญ



# Dropping Columns

ลบข้อมูลที่ไม่สำคัญ

1

แสดง Column  
ทั้งหมด

2

สร้าง Drop\_List  
และใส่ col ที่จะลบ

3

ทำการ Drop

# Dropping Columns

ลบข้อมูลที่ไม่สำคัญ

1

แสดง Column  
ทั้งหมด

2

สร้าง Drop\_List  
และใส่ col ที่จะลบ

3

ทำการ Drop

# All Column

```
▶ #Show Columns
print("All Columns")
for i, column in enumerate(data.columns):
    print(f"{i + 1}. {column}")
```

```
→ All Columns
1. Prospect ID
2. Lead Number
3. Lead Origin
4. Lead Source
5. Do Not Email
6. Do Not Call
7. Converted
8. TotalVisits
9. Total Time Spent on Website
10. Page Views Per Visit
11. Last Activity
12. Country
13. Specialization
14. How did you hear about X Education
15. What is your current occupation
16. What matters most to you in choosing a course
17. Search
18. Magazine
19. Newspaper Article
20. X Education Forums
21. Newspaper
22. Digital Advertisement
23. Through Recommendations
24. Receive More Updates About Our Courses
25. Tags
26. Lead Quality
27. Update me on Supply Chain Content
28. Get updates on DM Content
29. Lead Profile
30. City
31. Asymmetrique Activity Index
32. Asymmetrique Profile Index
33. Asymmetrique Activity Score
34. Asymmetrique Profile Score
35. I agree to pay the amount through cheque
36. A free copy of Mastering The Interview
37. Last Notable Activity
```

→ All Columns

1. Prospect ID
2. Lead Number
3. Lead Origin
4. Lead Source
5. Do Not Email
6. Do Not Call
7. Converted
8. TotalVisits
9. Total Time Spent on Website
10. Page Views Per Visit
11. Last Activity
12. Country
13. Specialization
14. How did you hear about X Education
15. What is your current occupation
16. What matters most to you in choosing a course
17. Search
18. Magazine
19. Newspaper Article
20. X Education Forums
21. Newspaper
22. Digital Advertisement
23. Through Recommendations
24. Receive More Updates About Our Courses
25. Tags
26. Lead Quality
27. Update me on Supply Chain Content
28. Get updates on DM Content
29. Lead Profile
30. City
31. Asymmetrique Activity Index
32. Asymmetrique Profile Index
33. Asymmetrique Activity Score
34. Asymmetrique Profile Score
35. I agree to pay the amount through cheque
36. A free copy of Mastering The Interview
37. Last Notable Activity

# Dropping Columns

ลบข้อมูลที่ไม่สำคัญ

1

แสดง Column  
ทั้งหมด

2

สร้าง Drop\_List  
และใส่ col ที่จะลบ

# Dropping Columns

ลบข้อมูลที่ไม่สำคัญ

1

แสดง Column  
ทั้งหมด

2

สร้าง Drop\_List  
และใส่ col ที่จะลบ

All Columns

- 1. Prospect ID
- 2. Lead Number
- 3. Lead Origin
- 4. Lead Source
- 5. Do Not Email
- 6. Do Not Call
- 7. Converted
- 8. TotalVisits
- 9. Total Time Spent on Website
- 10. Page Views Per Visit
- 11. Last Activity
- 12. Country
- 13. Specialization
- 14. How did you hear about X Education
- 15. What is your current occupation
- 16. What matters most to you in choosing a course
- 17. Search
- 18. Magazine
- 19. Newspaper Article
- 20. X Education Forums
- 21. Newspaper
- 22. Digital Advertisement
- 23. Through Recommendations
- 24. Receive More Updates About Our Courses
- 25. Tags
- 26. Lead Quality
- 27. Update me on Supply Chain Content
- 28. Get updates on DM Content
- 29. Lead Profile
- 30. City
- 31. Asymmetrique Activity Index
- 32. Asymmetrique Profile Index
- 33. Asymmetrique Activity Score
- 34. Asymmetrique Profile Score
- 35. I agree to pay the amount through cheque
- 36. A free copy of Mastering The Interview
- 37. Last Notable Activity



# Drop List

```
[ ] #Drop List
drop_list = ['How did you hear about X Education',
             'Lead Profile','Asymmetrique Activity Index',
             'Asymmetrique Activity Score',
             'Asymmetrique Profile Index',
             'Asymmetrique Profile Score',
             'Lead Number',
             'What matters most to you in choosing a course',
             'Search',
             'Magazine',
             'Newspaper Article',
             'X Education Forums',
             'Newspaper',
             'Digital Advertisement',
             'Through Recommendations',
             'Receive More Updates About Our Courses',
             'Update me on Supply Chain Content',
             'Get updates on DM Content',
             'I agree to pay the amount through cheque',
             'A free copy of Mastering The Interview',
             'Country']
```

# Dropping Columns

ลบข้อมูลที่ไม่สำคัญ

1

แสดง Column  
ทั้งหมด

2

สร้าง Drop\_List  
และใส่ col ที่จะลบ

3

ทำการ Drop

# Dropping Columns

ลบข้อมูลที่ไม่สำคัญ

1

แสดง Column  
ทั้งหมด

2

สร้าง Drop\_List  
และใส่ col ที่จะลบ

3

ทำการ Drop

# Dropping the Columns

หลังจาก drop\_list ให้ลบคอลัมน์ที่ระบุออกจาก DataFrame โดยใช้เมธอด .drop()

 #Dropping the Columns

```
data = data.drop(drop_list, axis = 1)
```



# Display the DataFrame

แสดงผลของ DataFrame หลังจากที่ลบคอลัมน์ที่ไม่ต้องการแล้ว

[ ] #Display the DataFrame  
data.head()

Prospect ID	Lead Origin	Lead Source	Do Not Email	Do Not Call	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Last Activity	Specialization	What is your current occupation	Tags	Lead Quality	City	Last Notable Activity
7927b2df-8bba-4d29-b9a2-b6e0beafe620	API	Olark Chat	No	No	0	0.0	0	0.0	Page Visited on Website	Select	Unemployed	Interested in other courses	Low in Relevance	Select	Modified
2a272436-5132-4136-86fa-dcc88c88f482	API	Organic Search	No	No	0	5.0	674	2.5	Email Opened	Select	Unemployed	Ringing	NaN	Select	Email Opened
8cc8c611-a219-4f35-ad23-fdd2656bd8a	Landing Page Submission	Direct Traffic	No	No	1	2.0	1532	2.0	Email Opened	Business Administration	Student	Will revert after reading the email	Might be	Mumbai	Email Opened
0cc2df48-7cf4-4e39-9de9-19797f9b38cc	Landing Page Submission	Direct Traffic	No	No	0	1.0	305	1.0	Unreachable	Media and Advertising	Unemployed	Ringing	Not Sure	Mumbai	Modified
3256f628-e534-4826-9d63-4a8b88782852	Landing Page Submission	Google	No	No	1	2.0	1428	1.0	Converted to Lead	Select	Unemployed	Will revert after reading the email	Might be	Mumbai	Modified



# Checking Duplicate Values in a Dataset

```
[ ] #Checking for Duplicate Values in a Dataset  
sum(data.duplicated(subset = 'Prospect ID')) == 0
```

→ True

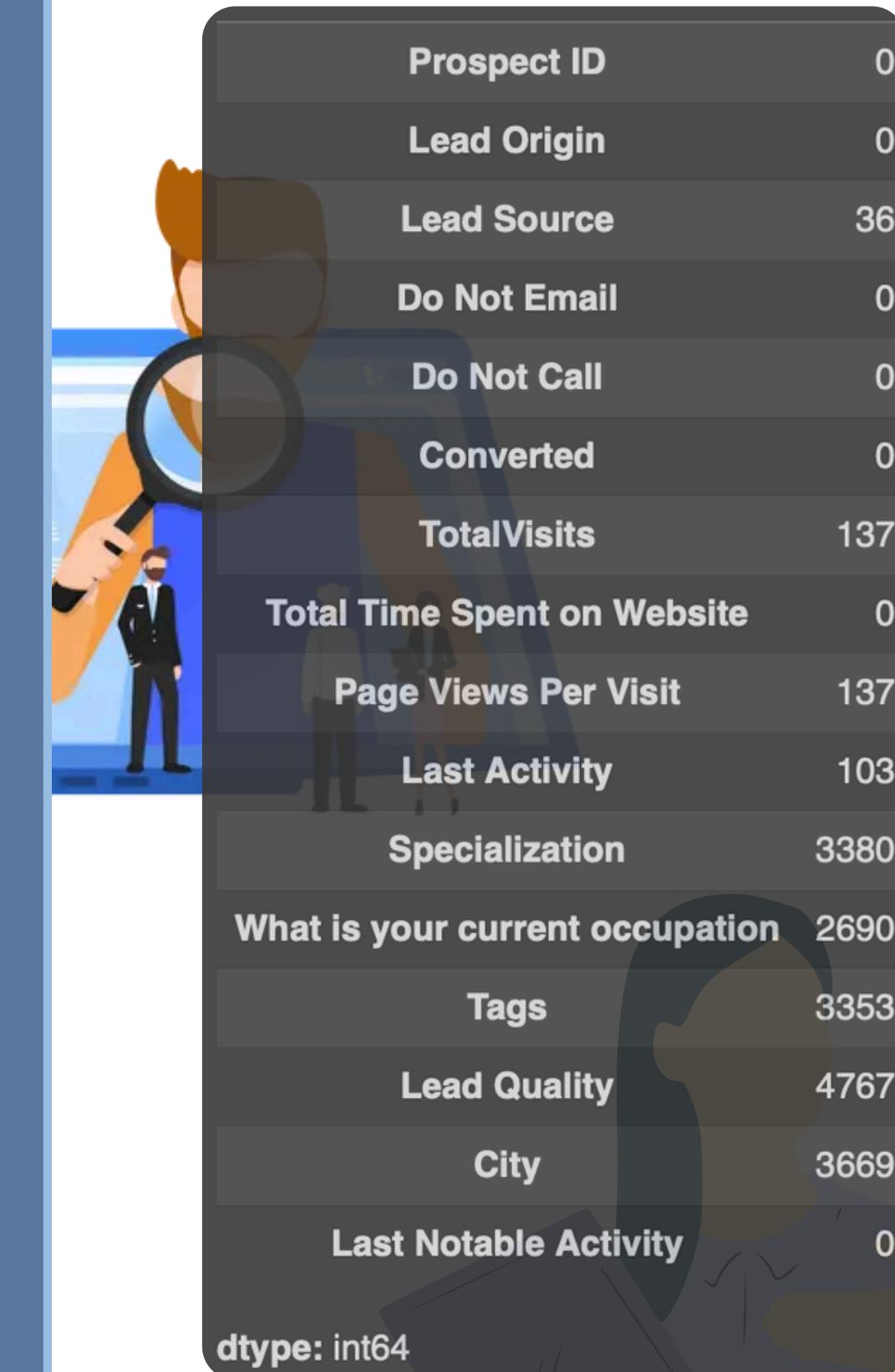
- True : ไม่มีแผลซ้ำกันในชุดข้อมูล
- False : มีแผลซ้ำกันในชุดข้อมูล

# Checking Null Values in a Dataset



```
#Check Null Values in a Dataset  
data = data.replace('Select', np.nan)  
data.isna().sum()
```

- จัดการข้อมูล Select
- ให้ np.nan (NaN) เข้าไปแทนค่า Select ผ่านฟังก์ชัน replace()
- ตรวจสอบค่าที่หายไปด้วย .isna() และนับค่าที่หายไปผ่าน .sum()



Prospect ID	0
Lead Origin	0
Lead Source	36
Do Not Email	0
Do Not Call	0
Converted	0
TotalVisits	137
Total Time Spent on Website	0
Page Views Per Visit	137
Last Activity	103
Specialization	3380
What is your current occupation	2690
Tags	3353
Lead Quality	4767
City	3669
Last Notable Activity	0
dtype: int64	





# Filling in Missing Value

# Filling in Missing Value

```


    #Filling in Missing Value
    for col in data.columns:
        if data[col].dtypes == 'int64' or data[col].dtypes == 'float64':
            data[col].fillna(data[col].mean(), inplace=True)
        else:
            data[col].fillna(data[col].mode()[0], inplace=True)

    data.isna().sum()

```

## DataFrame

	Prospect ID	Lead Origin	Lead Source	Do Not Email	Do Not Call	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Last Activity	Specialization	What is your current occupation	Tags	Lead Quality	City	Last Notable Activity
0	7927b2df-8bba-4d29-b9a2-b6e0beafe620	API	Olark Chat	No	No	0	0.0	0	0.0	Page Visited on Website	Select	Unemployed	Interested in other courses	Low in Relevance	Select	Modified
1	2a272436-5132-4136-86fa-dcc88c88f482	API	Organic Search	No	No	0	5.0	674	2.5	Email Opened	Select	Unemployed	Ringing	NaN	Select	Email Opened
2	8cc8c611-a219-4f35-ad23-fdfd2656bd8a	Landing Page Submission	Direct Traffic	No	No	1	2.0	1532	2.0	Email Opened	Business Administration	Student	Will revert after reading the email	Might be Mumbai	Email Opened	
3	0cc2df48-7cf4-4e39-9de9-19797f9b38cc	Landing Page Submission	Direct Traffic	No	No	0	1.0	305	1.0	Unreachable	Media and Advertising	Unemployed	Ringing	Not Sure	Mumbai	Modified
4	3256f628-e534-4826-9d63-4a8b88782852	Landing Page Submission	Google	No	No	1	2.0	1428	1.0	Converted to Lead	Select	Unemployed	Will revert after reading the email	Might be Mumbai	Modified	

# Filling in Missing Value

```
#Filling in Missing Value
for col in data.columns:
    if data[col].dtypes == 'int64' or data[col].dtypes == 'float64':
        data[col].fillna(data[col].mean(), inplace=True)
    else:
        data[col].fillna(data[col].mode()[0], inplace=True)

data.isna().sum()
```



## DataFrame

	Prospect ID	Lead Origin	Lead Source	Do Not Email	Do Not Call	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Last Activity	Specialization	What is your current occupation	Tags	Lead Quality	City	Last Notable Activity
0	7927b2df-8bba-4d29-b9a2-b6e0beafe620	API	Olark Chat	No	No	0	0.0	0	0.0	Page Visited on Website	Select	Unemployed	Interested in other courses	Low in Relevance	Select	Modified
1	2a272436-5132-4136-86fa-dcc88c88f482	API	Organic Search	No	No	0	5.0	674	2.5	Email Opened	Select	Unemployed	Ringing	NaN	Select	Email Opened
2	8cc8c611-a219-4f35-ad23-fdfd2656bd8a	Landing Page Submission	Direct Traffic	No	No	1	2.0	1532	2.0	Email Opened	Business Administration	Student	Will revert after reading the email	Might be Mumbai	Email Opened	
3	0cc2df48-7cf4-4e39-9de9-19797f9b38cc	Landing Page Submission	Direct Traffic	No	No	0	1.0	305	1.0	Unreachable	Media and Advertising	Unemployed	Ringing	Not Sure	Mumbai	Modified
4	3256f628-e534-4826-9d63-4a8b88782852	Landing Page Submission	Google	No	No	1	2.0	1428	1.0	Converted to Lead	Select	Unemployed	Will revert after reading the email	Might be Mumbai	Modified	

# Filling in Missing Value

```

#Filling in Missing Value
for col in data.columns:
    if data[col].dtypes == 'int64' or data[col].dtypes == 'float64':
        data[col].fillna(data[col].mean(), inplace=True)
    else:
        data[col].fillna(data[col].mode()[0], inplace=True)

data.isna().sum()

```



## DataFrame

	Prospect ID	Lead Origin	Lead Source	Do Not Email	Do Not Call	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Last Activity	Specialization	What is your current occupation	Tags	Lead Quality	City	Last Notable Activity
0	7927b2df-8bba-4d29-b9a2-b6e0beafe620	API	Olark Chat	No	No	0	0.0	0	0.0	Page Visited on Website	Select	Unemployed	Interested in other courses	Low in Relevance	Select	Modified
1	2a272436-5132-4136-86fa-dcc88c88f482	API	Organic Search	No	No	0	5.0	674	2.5	Email Opened	Select	Unemployed	Ringing	NaN	Select	Email Opened
2	8cc8c611-a219-4f35-ad23-fdfd2656bd8a	Landing Page Submission	Direct Traffic	No	No	1	2.0	1532	2.0	Email Opened	Business Administration	Student	Will revert after reading the email	Might be Mumbai	Email Opened	
3	0cc2df48-7cf4-4e39-9de9-19797f9b38cc	Landing Page Submission	Direct Traffic	No	No	0	1.0	305	1.0	Unreachable	Media and Advertising	Unemployed	Ringing	Not Sure	Mumbai	Modified
4	3256f628-e534-4826-9d63-4a8b88782852	Landing Page Submission	Google	No	No	1	2.0	1428	1.0	Converted to Lead	Select	Unemployed	Will revert after reading the email	Might be Mumbai	Modified	

# Filling in Missing Value

```
#Filling in Missing Value
for col in data.columns:
    if data[col].dtypes == 'int64' or data[col].dtypes == 'float64':
        data[col].fillna(data[col].mean(), inplace=True)
    else:
        data[col].fillna(data[col].mode()[0], inplace=True)

data.isna().sum()
```



## DataFrame

	Prospect ID	Lead Origin	Lead Source	Do Not Email	Do Not Call	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Last Activity	Specialization	What is your current occupation	Tags	Lead Quality	City	Last Notable Activity
0	7927b2df-8bba-4d29-b9a2-b6e0beafe620	API	Olark Chat	No	No	0	0.0	0	0.0	Page Visited on Website	Select	Unemployed	Interested in other courses	Low in Relevance	Select	Modified
1	2a272436-5132-4136-86fa-dcc88c88f482	API	Organic Search	No	No	0	5.0	674	2.5	Email Opened	Select	Unemployed	Ringing	NaN	Select	Email Opened
2	8cc8c611-a219-4f35-ad23-fdfd2656bd8a	Landing Page Submission	Direct Traffic	No	No	1	2.0	1532	2.0	Email Opened	Business Administration	Student	Will revert after reading the email	Might be	Mumbai	Email Opened
3	0cc2df48-7cf4-4e39-9de9-19797f9b38cc	Landing Page Submission	Direct Traffic	No	No	0	1.0	305	1.0	Unreachable	Media and Advertising	Unemployed	Ringing	Not Sure	Mumbai	Modified
4	3256f628-e534-4826-9d63-4a8b88782852	Landing Page Submission	Google	No	No	1	2.0	1428	1.0	Converted to Lead	Select	Unemployed	Will revert after reading the email	Might be	Mumbai	Modified

# Filling in Missing Value

Prospect ID	0
Lead Origin	0
Lead Source	36
Do Not Email	0
Do Not Call	0
Converted	0
TotalVisits	137
Total Time Spent on Website	0
Page Views Per Visit	137
Last Activity	103
Specialization	3380
What is your current occupation	2690
Tags	3353
Lead Quality	4767
City	3669
Last Notable Activity	0
dtype: int64	

.fillna()

Prospect ID	0
Lead Origin	0
Lead Source	0
Do Not Email	0
Do Not Call	0
Converted	0
TotalVisits	0
Total Time Spent on Website	0
Page Views Per Visit	0
Last Activity	0
Specialization	0
What is your current occupation	0
Tags	0
Lead Quality	0
City	0
Last Notable Activity	0
dtype: int64	

# การปรับแต่งการแสดงผล และการแสดงข้อมูล



# — Data Display — — Customization —



#Data Display Customization

```
pd.set_option('display.max_rows', 20)
```

```
pd.set_option('display.max_columns', 100)
```

```
pd.options.display.max_rows = 100
```

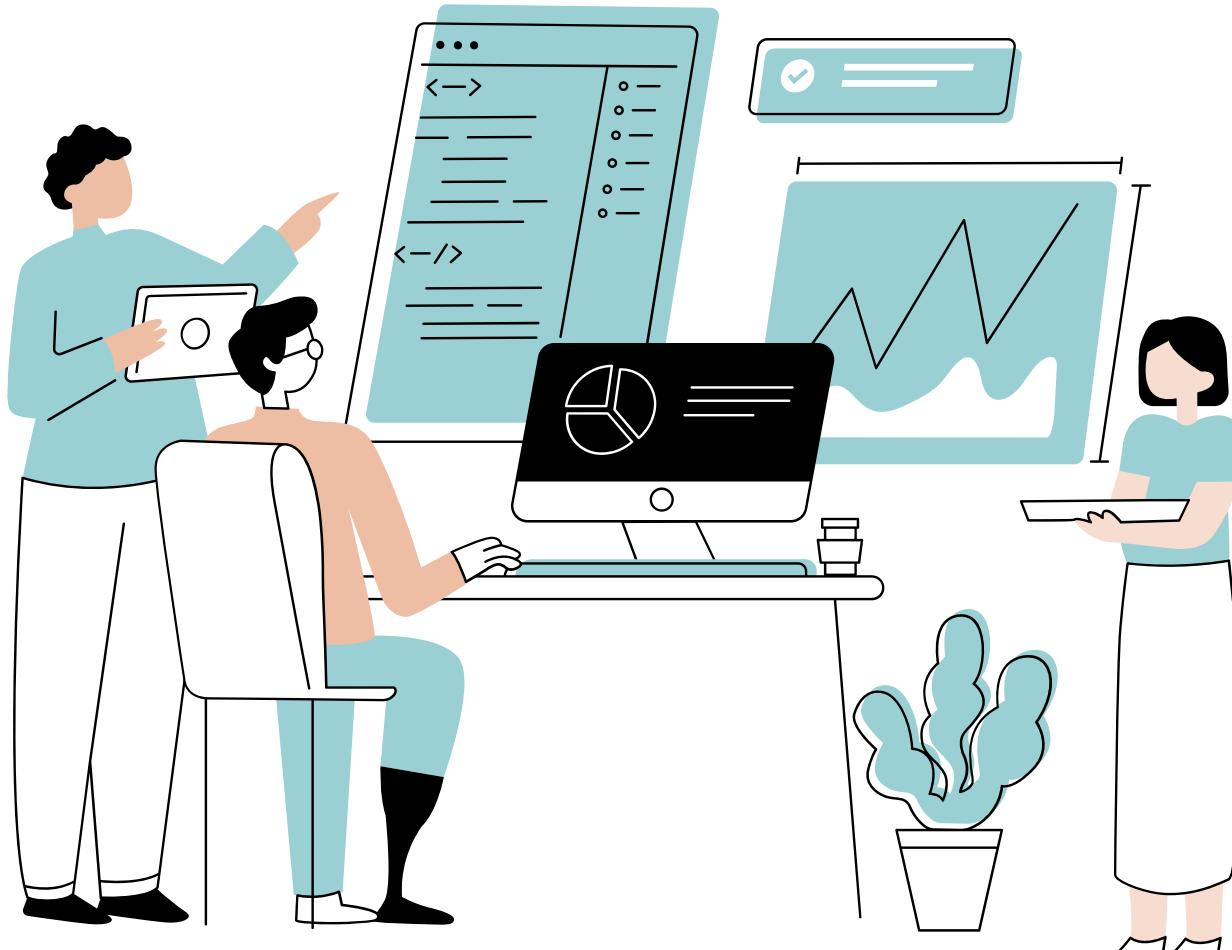


	Prospect ID	Lead Origin	Lead Source	Do Not Email	Do Not Call	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Last Activity	Specialization	What is your current occupation	Tags	Lead Quality	City	Last Notable Activity
0	7927b2df-8bba-4d29-b9a2-b6e0beafe620	API	Olark Chat	No	No	0	0.0	0	0.0	Page Visited on Website	Select	Unemployed	Interested in other courses	Low in Relevance	Select	Modified
1	2a272436-5132-4136-86fa-dcc88c88f482	API	Organic Search	No	No	0	5.0	674	2.5	Email Opened	Select	Unemployed	Ringing	NaN	Select	Email Opened
2	8cc8c611-a219-4f35-ad23-fdfd2656bd8a	Landing Page Submission	Direct Traffic	No	No	1	2.0	1532	2.0	Email Opened	Business Administration	Student	Will revert after reading the email	Might be	Mumbai	Email Opened
3	0cc2df48-7cf4-4e39-9de9-19797f9b38cc	Landing Page Submission	Direct Traffic	No	No	0	1.0	305	1.0	Unreachable	Media and Advertising	Unemployed	Ringing	Not Sure	Mumbai	Modified
4	3256f628-e534-4826-9d63-4a8b88782852	Landing Page Submission	Google	No	No	1	2.0	1428	1.0	Converted to Lead	Select	Unemployed	Will revert after reading the email	Might be	Mumbai	Modified



```
[ ] #Visualization  
from matplotlib.pyplot import xticks  
%matplotlib inline
```

ตั้งค่า Matplotlib ช่วยให้เราปรับแต่งลักษณะของกราฟได้ในโค้ดนี้จะใช้ %matplotlib inline เป็น magic command เพื่อให้กราฟแสดงผลอยู่ภายใน notebook



# Visualization

ด้วยปัจจุบัน

```
plt.rcParams['figure.figsize'] = (10, 6) # ขนาดของกราฟ (กว้าง x สูง)  
plt.rcParams['axes.titlesize'] = 14 # ขนาดของข้อความชื่อแกน  
plt.rcParams['axes.labelsize'] = 12 # ขนาดของข้อความ label ของแกน  
plt.rcParams['xtick.labelsize'] = 10 # ขนาดของข้อความ labels บนแกน x  
plt.rcParams['ytick.labelsize'] = 10 # ขนาดของข้อความ labels บนแกน y  
plt.rcParams['legend.fontsize'] = 12 # ขนาดของข้อความ legend  
plt.rcParams['grid.color'] = 'gray' # สีของกริด  
plt.rcParams['grid.linestyle'] = '--' # ลักษณะของเส้นกริด
```

ปรับแต่งและแสดงข้อมูล



# Converted



Leads



Customer

# Converted

```
▶ data['Converted'].value_counts()
```

## Output

	count
<b>Converted</b>	
0	5679
1	3561
dtype: int64	



9240

Leads

# Converted

# 38.54 %



Leads



Customer

```
▶ converted = round(sum(data['Converted']) / len(data['Converted'])*100, 2)  
print(converted, '%')  
→ 38.54 %
```



# Lead Origin

ที่มาของลูกค้าที่มีความสนใจต่อสินค้า

# Lead Origin

```
[ ] data['Lead Origin'].value_counts()
```

count

Lead Origin	count
Landing Page Submission	4886
API	3580
Lead Add Form	718
Lead Import	55
Quick Add Form	1

dtype: int64

# Lead Origin

```
[ ] data['Lead Origin'].value_counts()
```

	count
Lead Origin	
Landing Page Submission	4886
API	3580
Lead Add Form	718
Lead Import	55
Quick Add Form	1

dtype: int64

Dropping

```
[ ] # ลบแควทั้งหมดของ Quick Add Form  
data.drop(data.index[data['Lead Origin'] == 'Quick Add Form'], inplace=True)
```

จากตรงนี้พบว่า Quick Add Form มีจำนวนข้อมูลกี่น้อยมาก  
ดังนั้นจะทำการลบชุดข้อมูลนั้นออก

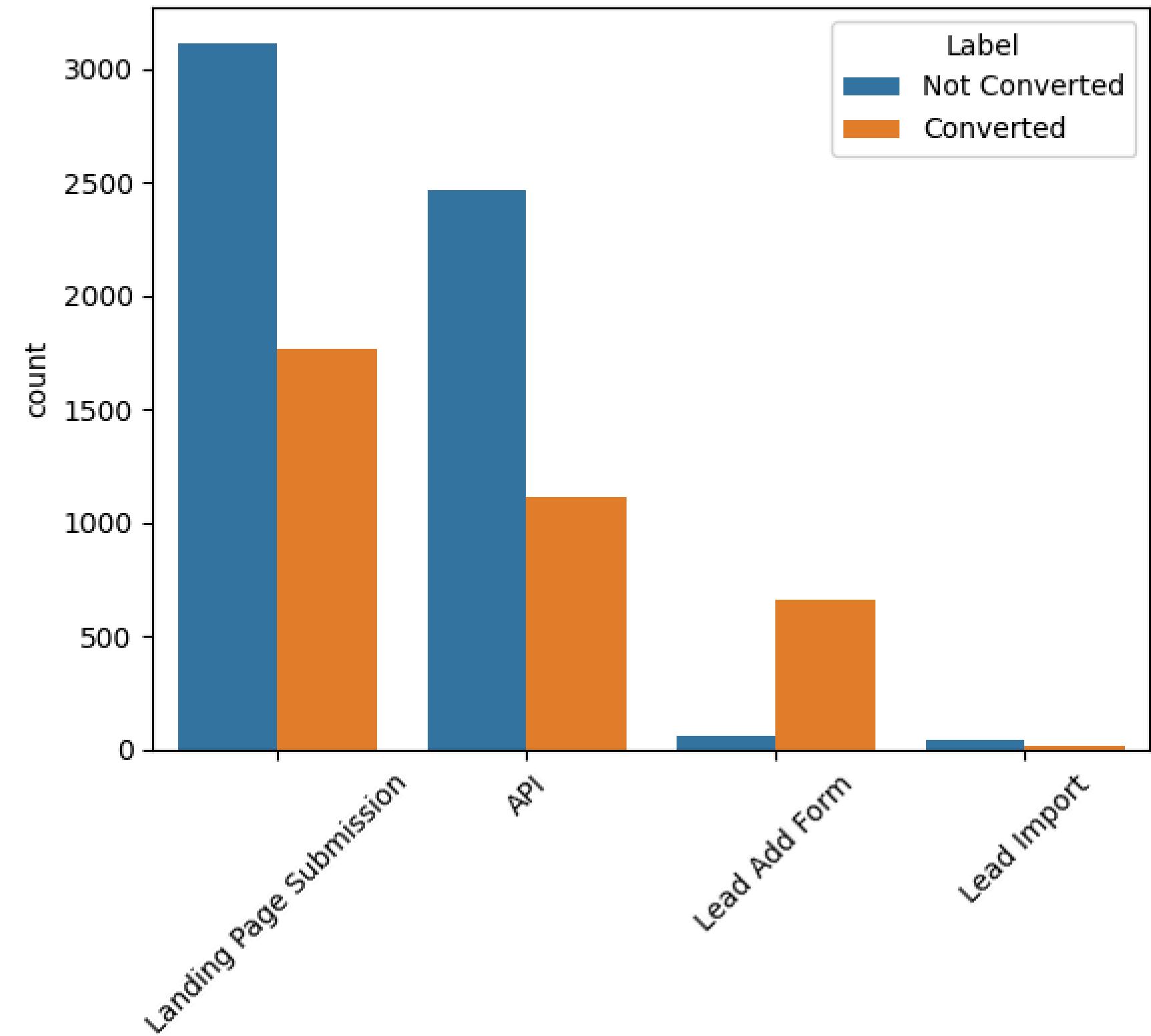
# Lead Origin



```
[ ] # สร้างคอลัมน์ใหม่ที่เปลี่ยนแปลงค่า 0 และ 1 ให้เป็น Converted และ Not Converted เพื่อความเข้าใจในการอ่านกราฟ
data['Label'] = data['Converted'].map({1: 'Converted', 0: 'Not Converted'})

# สร้างกราฟ Lead Origin โดยแยก Lead ที่ถูก Converted จากสี
sns.countplot(x="Lead Origin", hue="Label", data=data[['Lead Origin']].value_counts().index)
plt.xticks(rotation=45)
plt.show()
```

กลับจากจัดข้อมูลเรียบร้อยแล้วให้แสดงจำนวนลูกค้าที่สนใจสินค้าโดยแบ่งแยกตามแหล่งที่มา



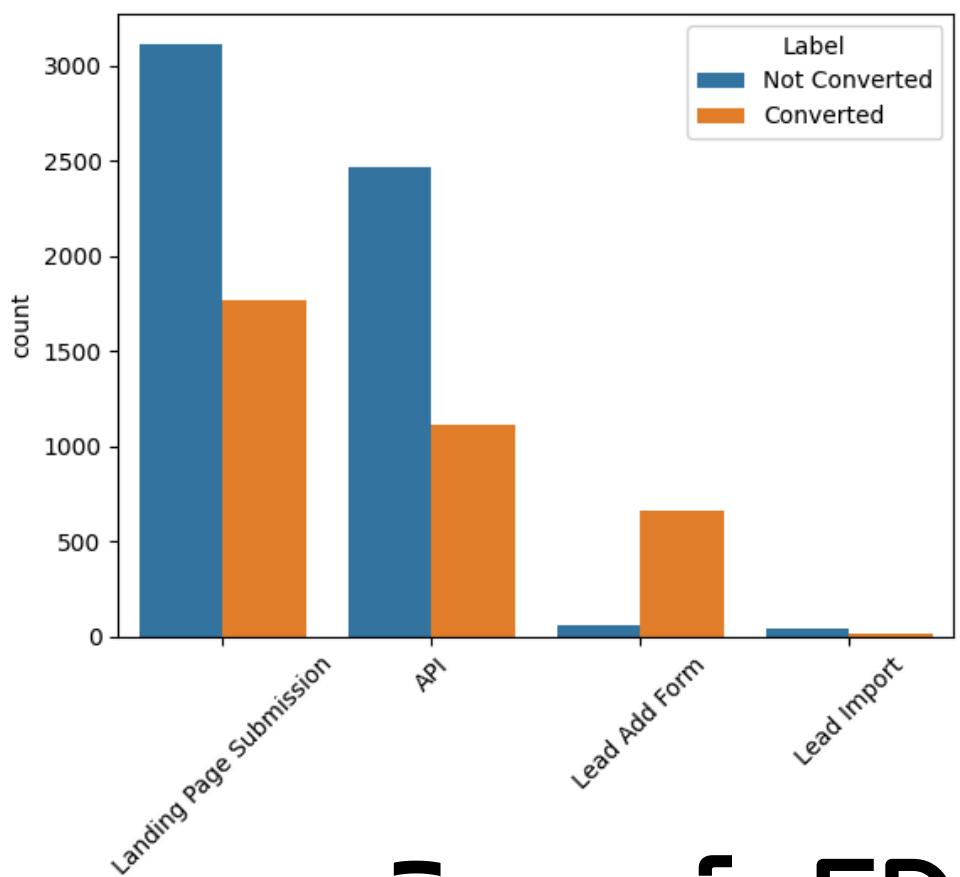
# Lead Origin

```
# คำนวณอัตราการเปลี่ยนแปลงสถานะของลูกค้ากับแหล่งที่มาของลูกค้า จากนั้นแสดงผลออกมานเป็น DataFrame
d = {}
for val in data['Lead Origin'].unique():
    a = data[data['Lead Origin'] == val]['Converted'].sum()
    b = data[data['Lead Origin'] == val]['Converted'].count()
    d[val] = [a, b, round(a/b*100, 2)]
pd.DataFrame.from_dict(d, orient='index').rename(columns = {0: 'Converted', 1: 'Leads', 2: 'Conversion Ratio'}).sort_values(by=['Conversion Ratio'], ascending=False)
```

	Converted	Leads	Conversion Ratio
<b>Lead Add Form</b>	664	718	92.48
<b>Landing Page Submission</b>	1768	4886	36.19
<b>API</b>	1115	3580	31.15
<b>Lead Import</b>	13	55	23.64

อัตราการเปลี่ยนแปลงสถานะของลูกค้ากับแหล่งที่มาของลูกค้า

# Output



## การวิเคราะห์ EDA

# Lead Source

แสดงข้อมูลช่องทางทั้งหมดที่นำลูกค้าที่มีความสนใจเข้ามาหาสินค้า

การวิเคราะห์ EDA

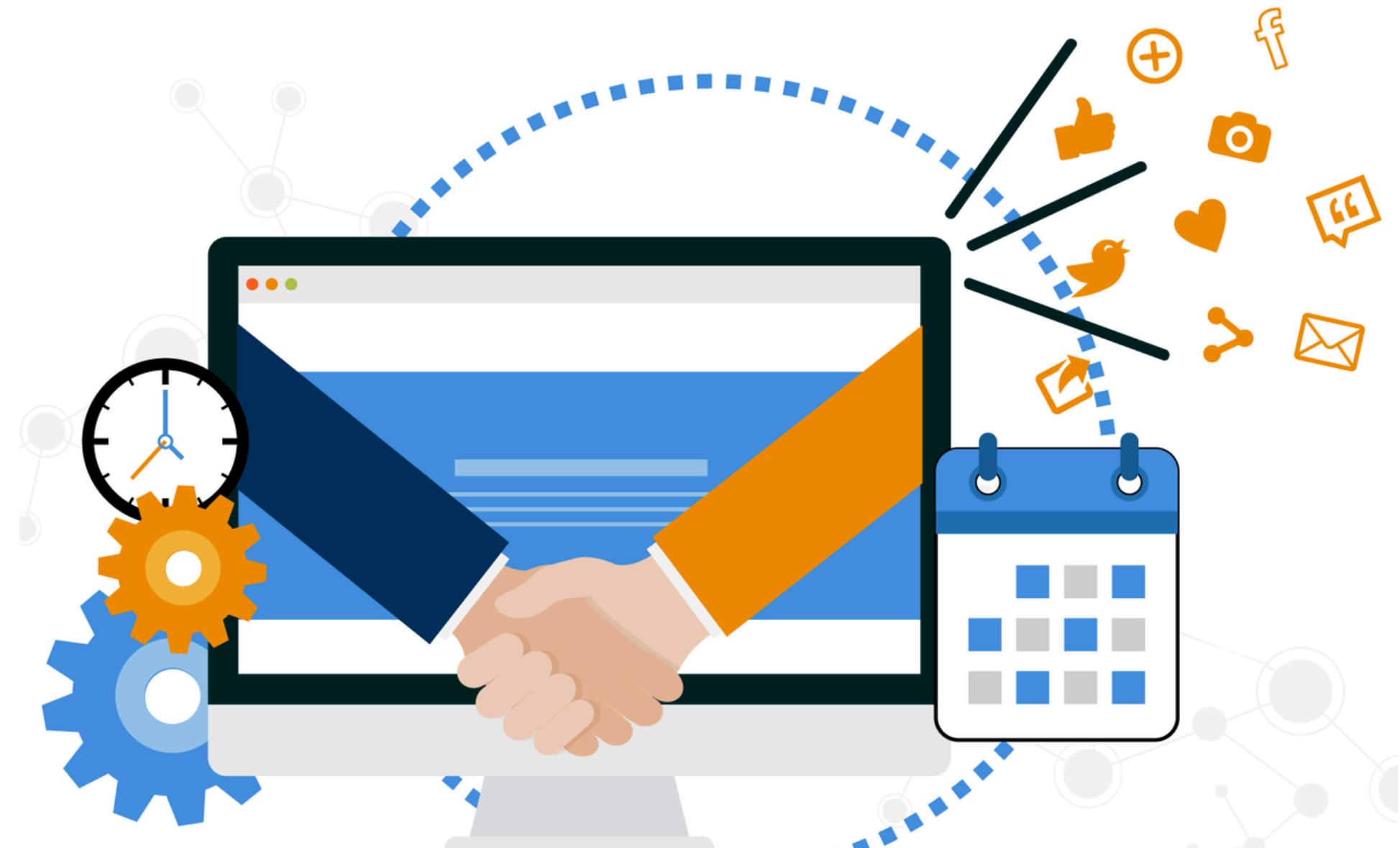
# Lead Source

แสดงข้อมูลช่องทางทั้งหมดที่นำลูกค้าที่มีความสนใจเข้ามาหาสินค้า

```
data['Lead Source'].value_counts()
```

Lead Source	
Google	2903
Direct Traffic	2543
Olark Chat	1755
Organic Search	1154
Reference	534
Welingak Website	142
Referral Sites	125
Facebook	55
bing	6
google	5
Click2call	4
Press_Release	2
Social Media	2
Live Chat	2
youtubechannel	1
testone	1
Pay per Click Ads	1
welearnblog_Home	1
WeLearn	1
blog	1
NC_EDM	1

dtype: int64



# Lead Source

แสดงข้อมูลช่องทางทั้งหมดที่นำลูกค้าที่มีความสนใจเข้ามาหาสินค้า

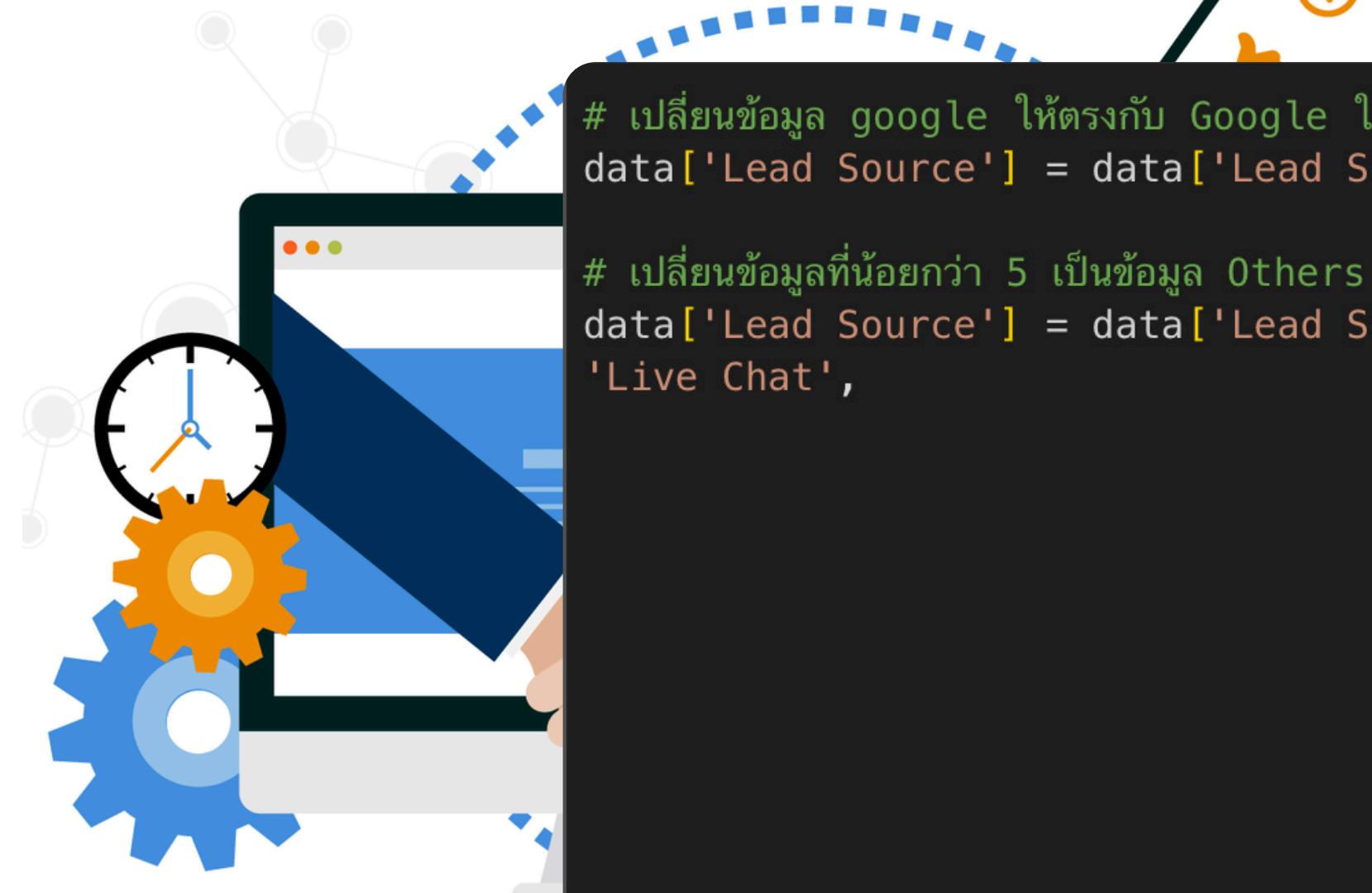
```
data['Lead Source'].value_counts()
```

Lead Source	
Google	2903
Direct Traffic	2543
Dark Chat	1755
Organic Search	1154
Reference	534
Welingak Website	142
Referral Sites	125
Facebook	55
bina	6
google	5
Click2call	4
Press_Release	2
Social Media	2
Live Chat	2
youtubechannel	1
testone	1
Pay per Click Ads	1
welearnblog_Home	1
WeLearn	1
blog	1
NC_EDM	1

dtype: int64

จากข้อมูลพบว่ามี Google อยู่ 2 ที่และยังมีอีกหลายชุดข้อมูลที่มีจำนวนน้อยเกินไป

ให้ทำการรวมข้อมูลก่อนไปแสดงผล

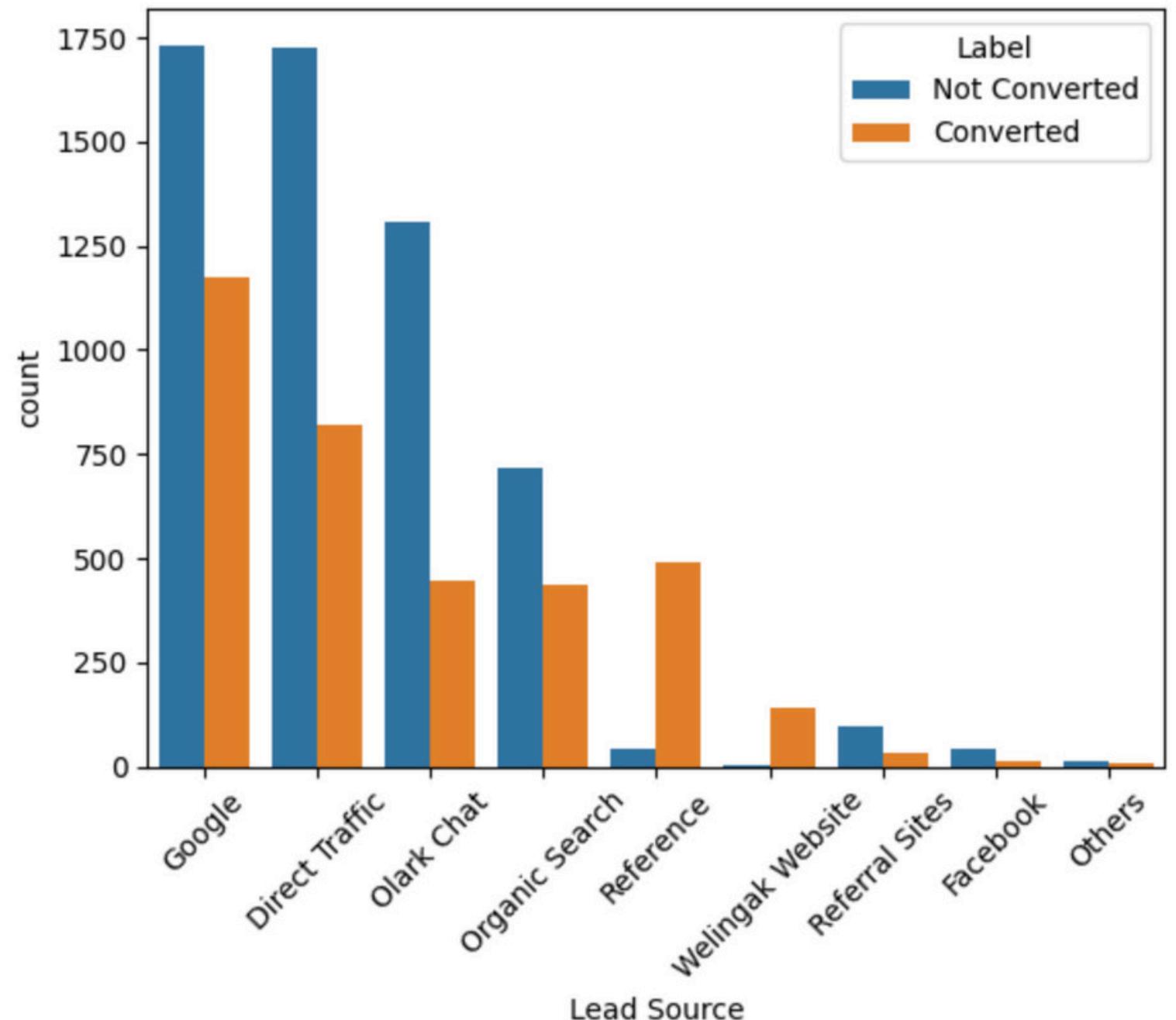


```
# เปลี่ยนข้อมูล google ให้ตรงกับ Google ในชุดข้อมูล
data['Lead Source'] = data['Lead Source'].replace(['google'], 'Google')

# เปลี่ยนข้อมูลที่น้อยกว่า 5 เป็นข้อมูล Others
data['Lead Source'] = data['Lead Source'].replace(['Click2call',
'Live Chat',
'NC_EDM',
'Pay per Click Ads',
'Press_Release',
'Social Media',
'WeLearn',
'bing',
'blog',
'testone',
'welearnblog_Home',
'youtubechannel'], 'Others')
```

```
sns.countplot(x="Lead Source", hue="Label", data=data, order=data['Lead Source'].value_counts().index  
plt.xticks(rotation=45)  
plt.show()
```

GROUP 20

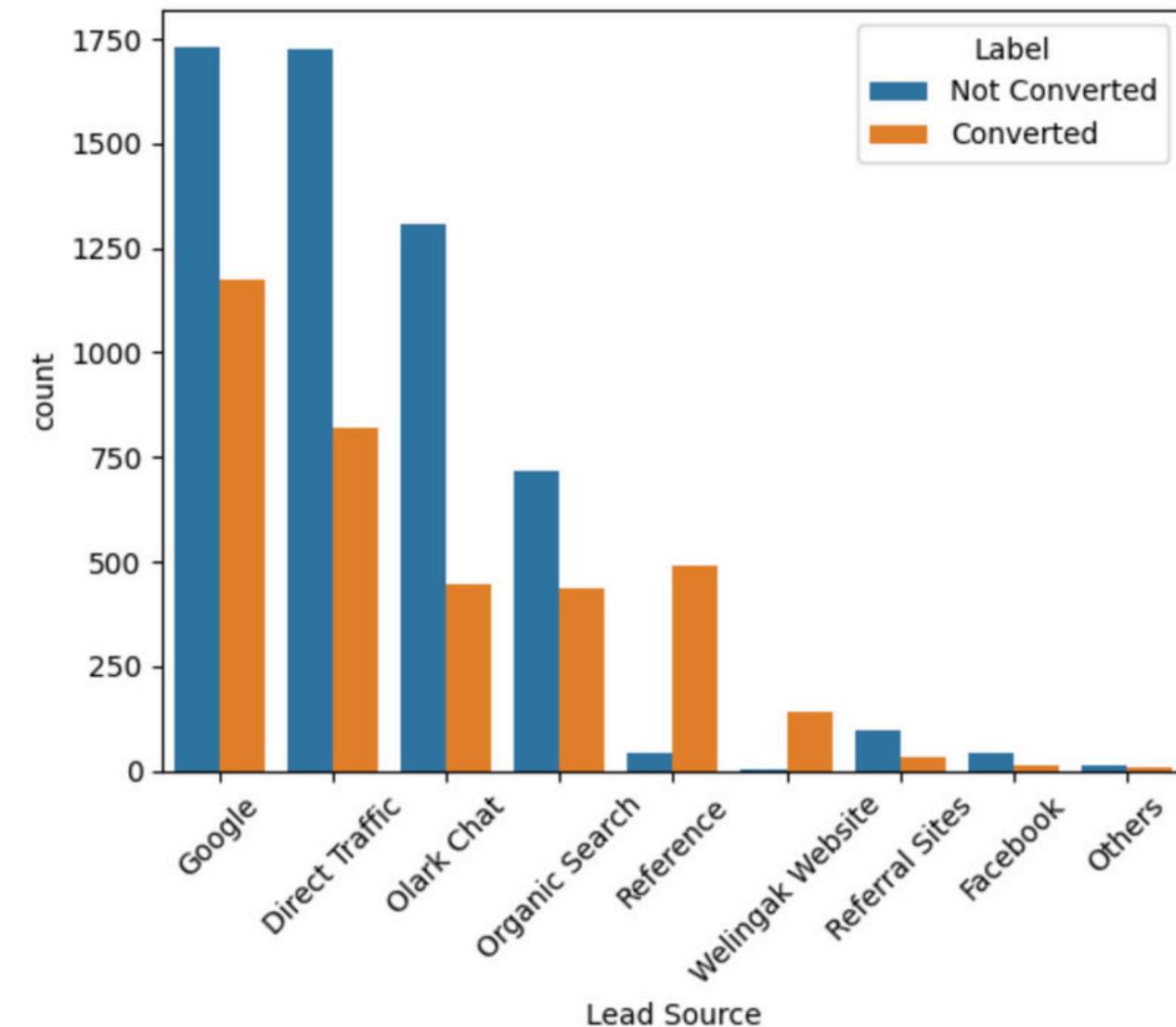


# Lead Source

## คำนวณอัตราการเปลี่ยนแปลงสถานะของลูกค้ากับช่องทางที่ลูกค้าเข้ามาหา

GROUP 20

```
# คำนวณอัตราการเปลี่ยนแปลงสถานะของลูกค้ากับช่องทางที่ลูกค้าเข้ามาหา จากนั้นแสดงผลออกมาเป็น DataFrame  
d = {}  
for val in data['Lead Source'].unique():  
    a = data[data['Lead Source'] == val]['Converted'].sum()  
    b = data[data['Lead Source'] == val]['Converted'].count()  
    d[val] = [a, b, round(a/b*100, 2)]  
pd.DataFrame.from_dict(d, orient='index').rename(columns = {0: 'Converted', 1: 'Leads',2: 'Conversion Ratio'}).sort_values(by=['Conversion Ratio'], ascending=False)
```



	Converted	Leads	Conversion Ratio
Welingak Website	140	142	98.59
Reference	490	534	91.76
Google	1175	2908	40.41
Others	9	23	39.13
Organic Search	436	1154	37.78
Direct Traffic	818	2543	32.17
Olark Chat	448	1755	25.53
Referral Sites	31	125	24.80
Facebook	13	55	23.64

# Lead Source

# Total Time Spent on Website

---

แสดงข้อมูลเวลาทั้งหมดที่ลูกค้าเข้าชมเว็บไซต์





แสดงค่าสถิติเบื้องต้น โดยใช้ .describe()

```
data['Total Time Spent on Website'].describe()
```

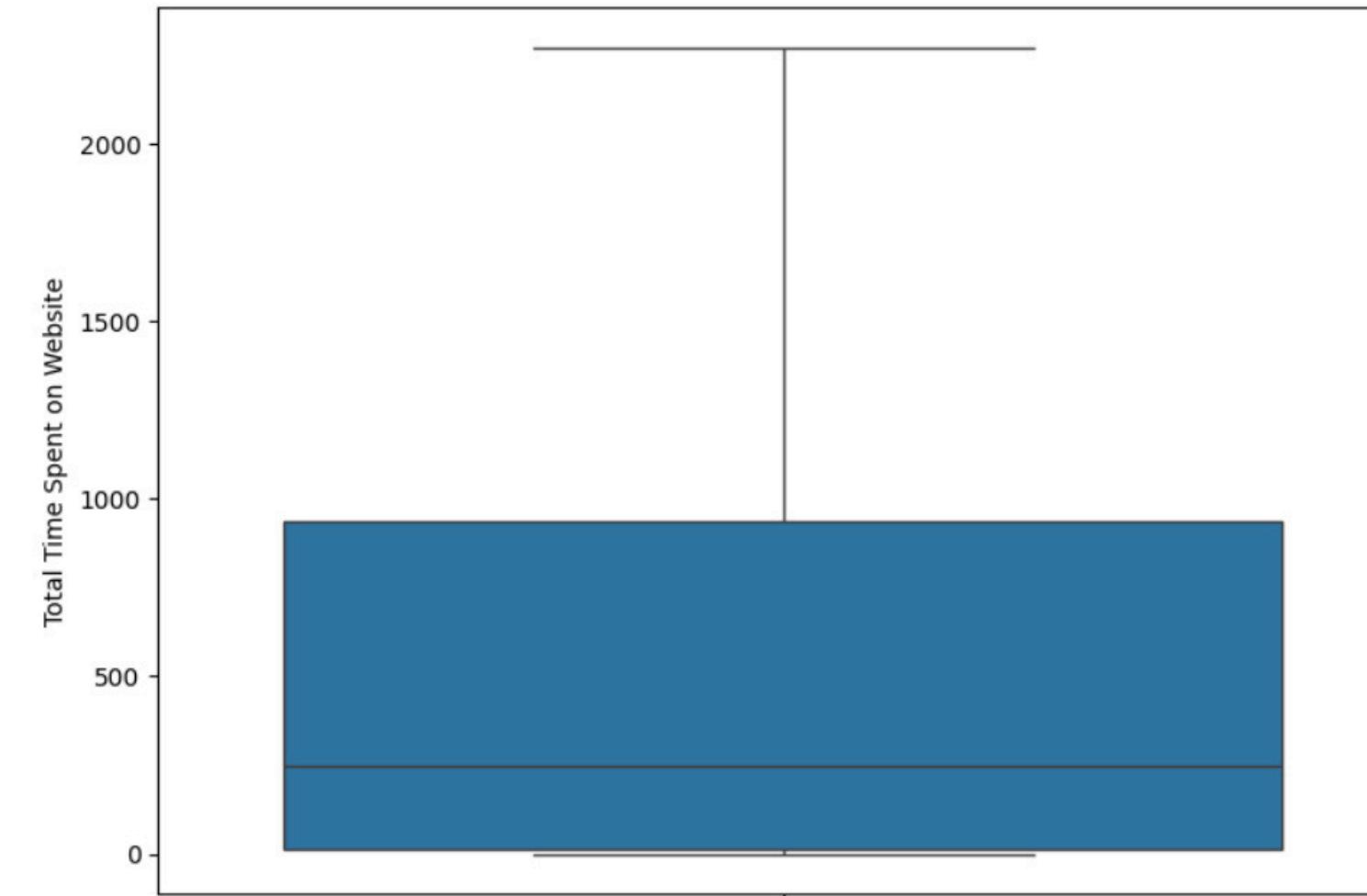
Total Time Spent on Website	
count	9239.000000
mean	487.511094
std	547.755682
min	0.000000
25%	12.000000
50%	248.000000
75%	936.000000
max	2272.000000

## Total Time Spent on Website

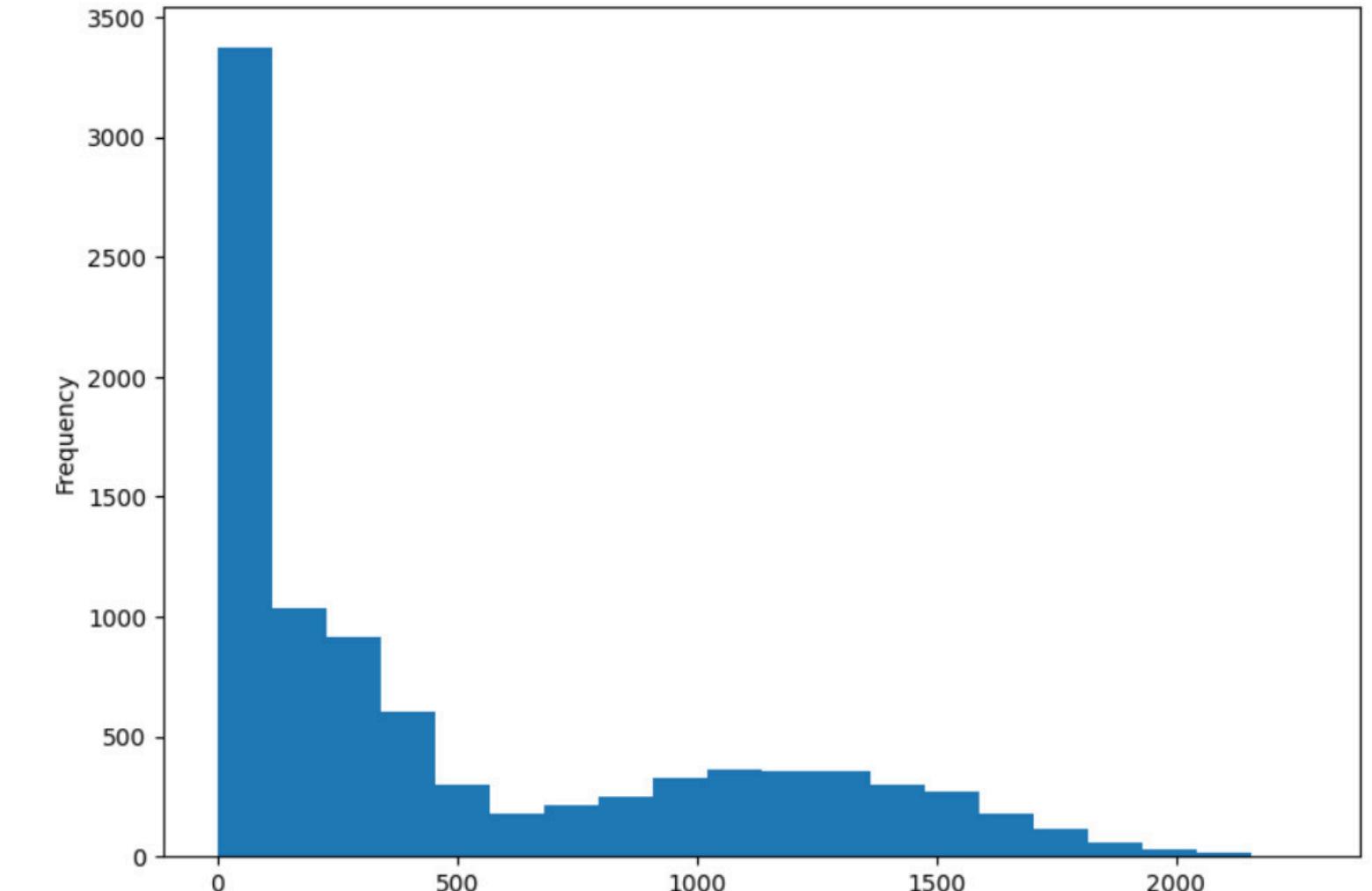
แสดงข้อมูลเวลาทั้งหมดที่ลูกค้าเข้าชมเว็บไซต์

สร้างกราฟ 2 ชนิดบน subplot เดียวกัน โดยแบ่งเป็น 2 คอลัมน์

```
# plotting boxplot and histogram
fig, axs = plt.subplots(1,2, figsize = (20,6.5)) #
sns.boxplot(data['Total Time Spent on Website'], ax = axs[0])
data['Total Time Spent on Website'].plot.hist(bins=20, ax = axs[1])
plt.show()
```



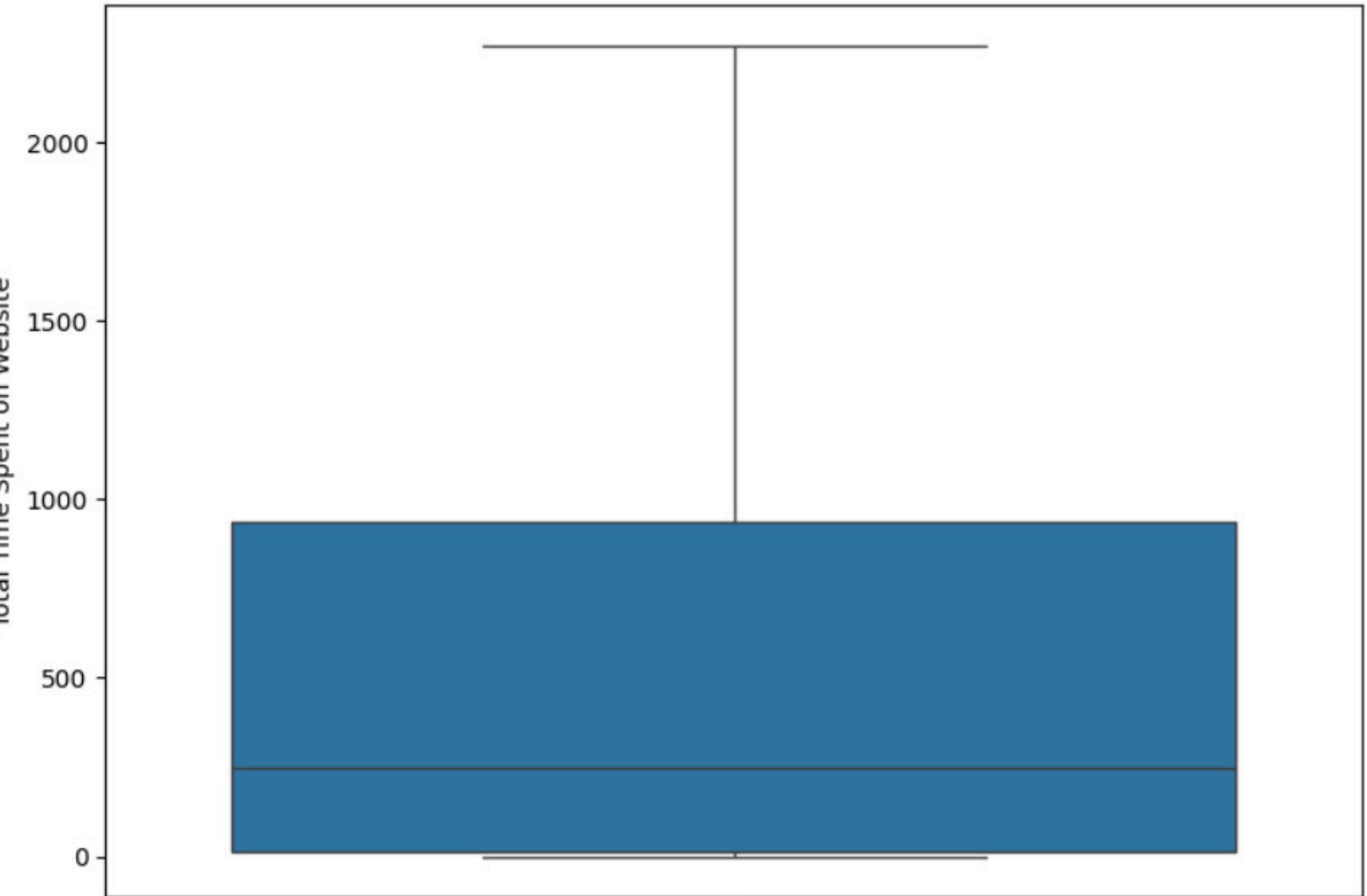
Box Plot



Histogram

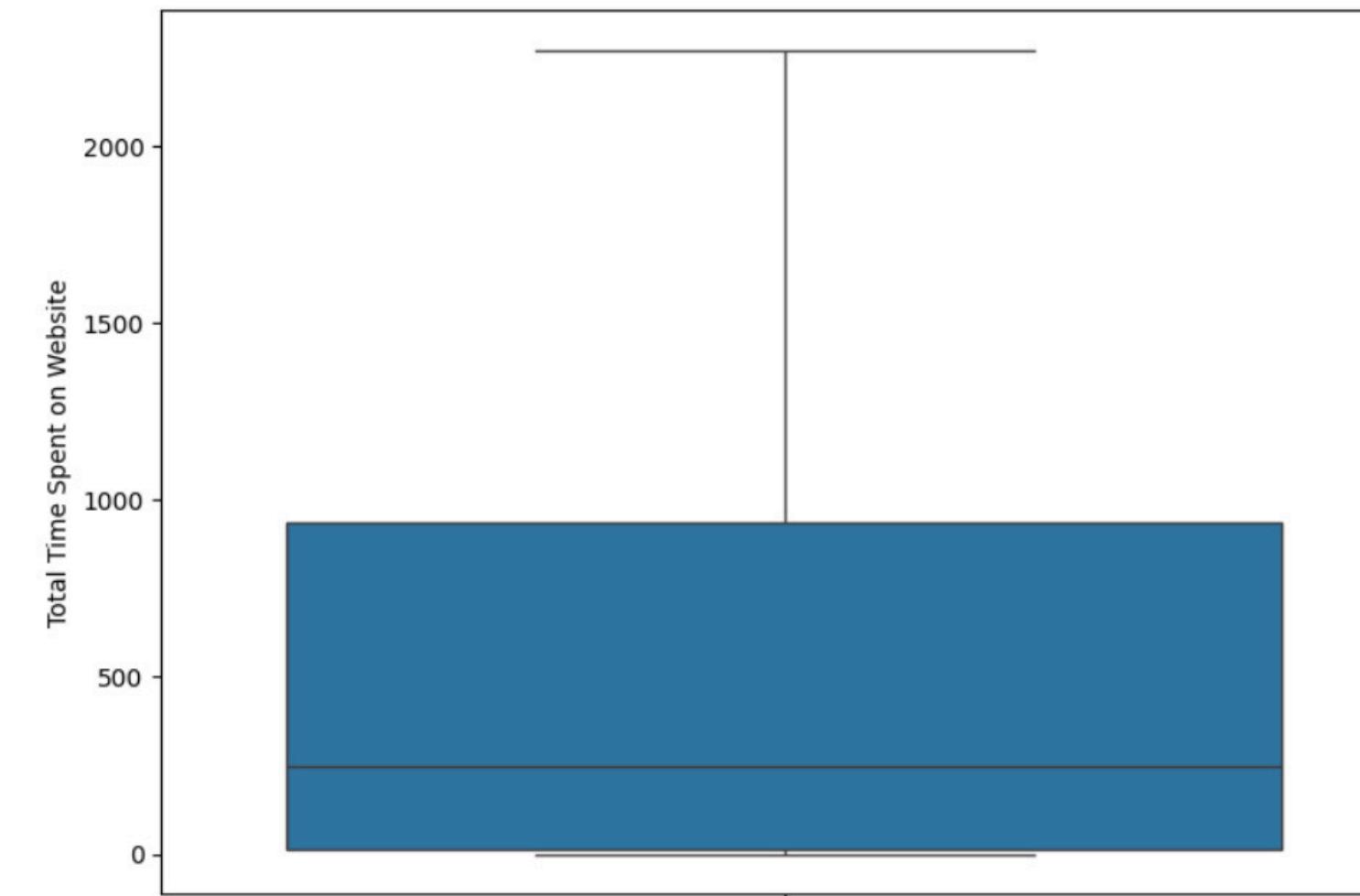
# Box Plot

- กราฟ Boxplot แสดงการกระจายตัวของข้อมูล “Total Time Spent on Website” (เวลาเฉลี่ยที่ผู้ใช้ใช้บนเว็บไซต์) ซึ่งช่วยให้เห็นค่ามัธยฐานและค่าผิดปกติได้ชัดเจน

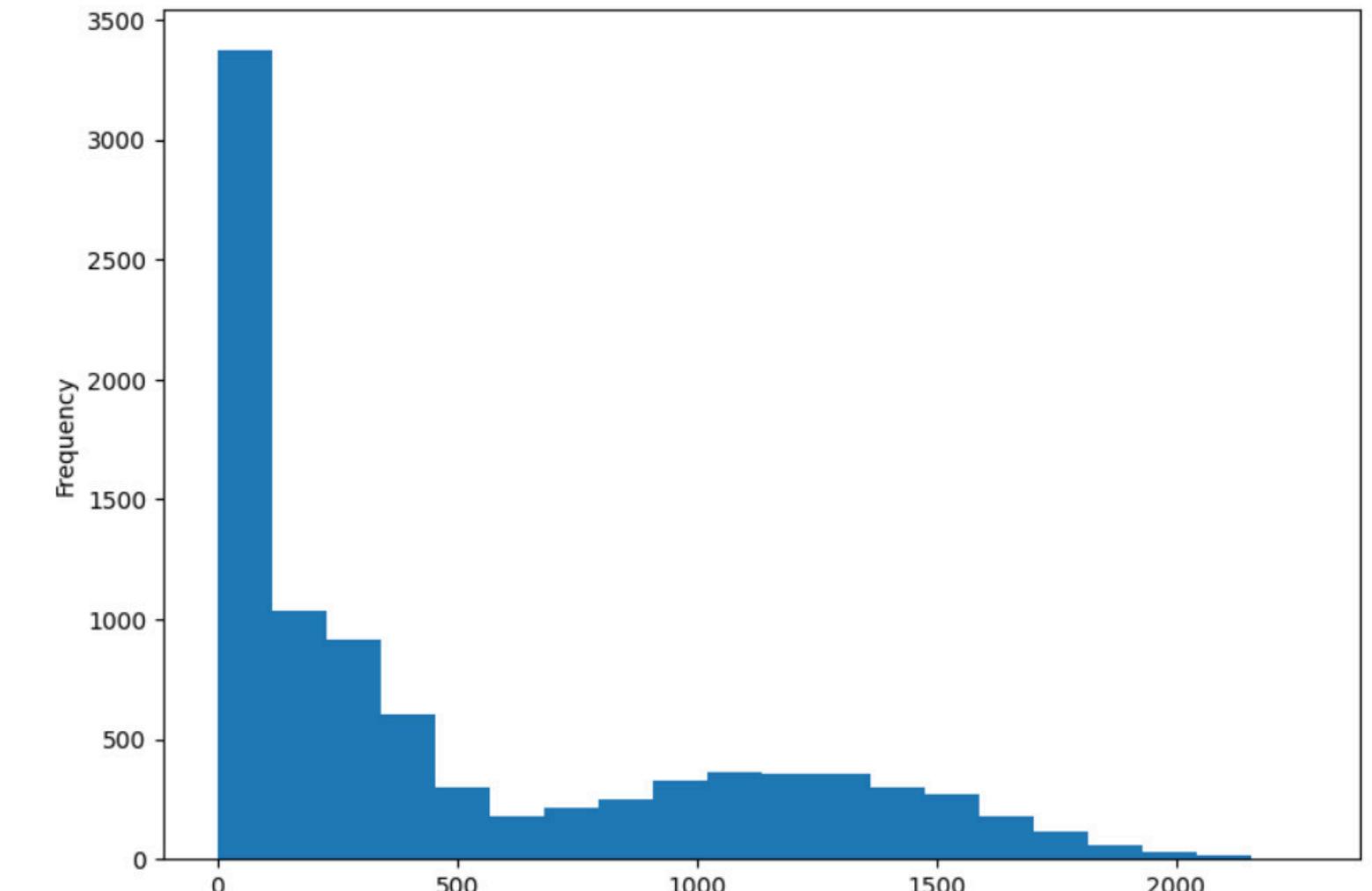


สร้างกราฟ 2 ชนิดบน subplot เดียวกัน โดยแบ่งเป็น 2 คอลัมน์

```
# plotting boxplot and histogram
fig, axs = plt.subplots(1,2, figsize = (20,6.5)) #
sns.boxplot(data['Total Time Spent on Website'], ax = axs[0])
data['Total Time Spent on Website'].plot.hist(bins=20, ax = axs[1])
plt.show()
```

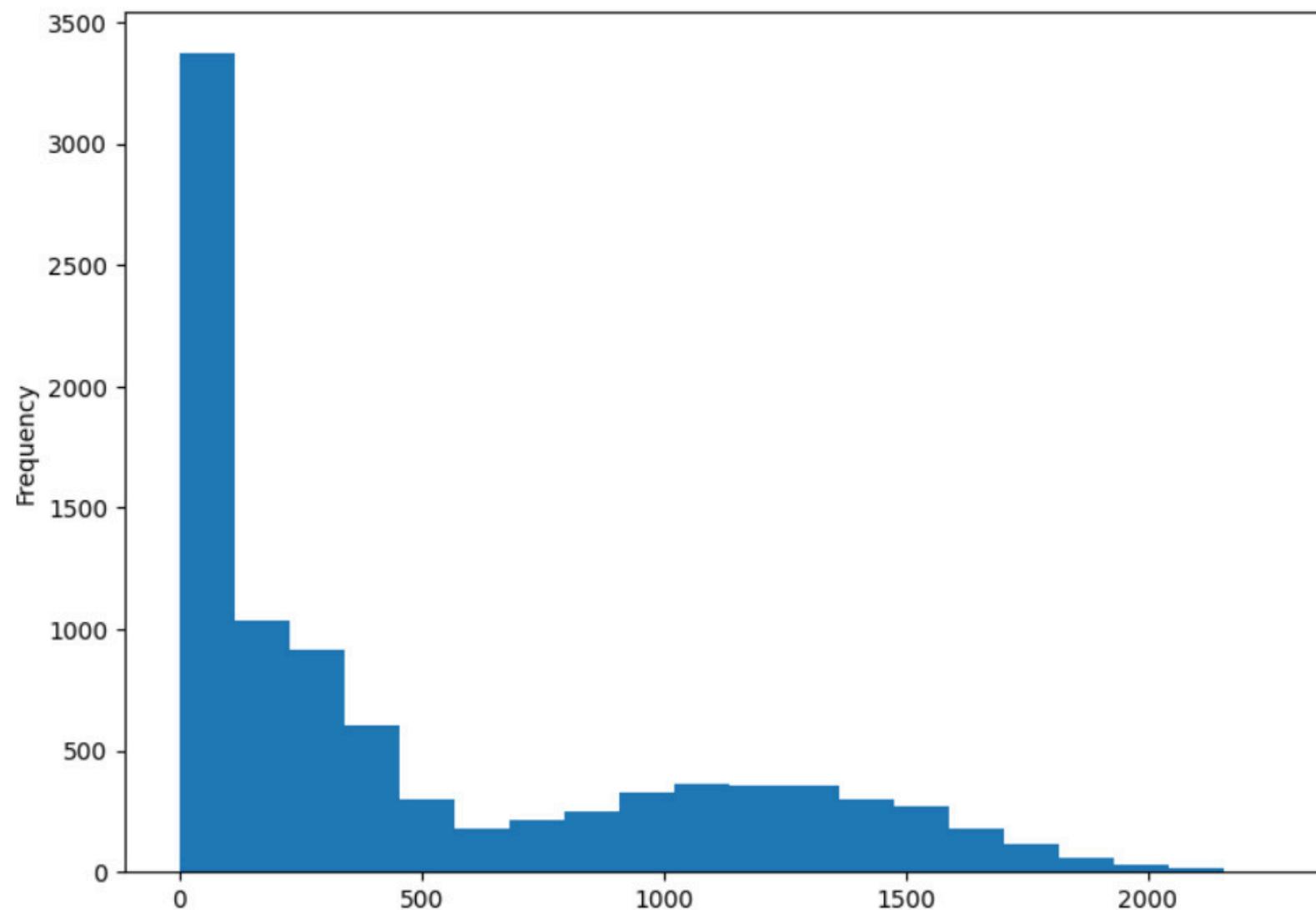


Box Plot



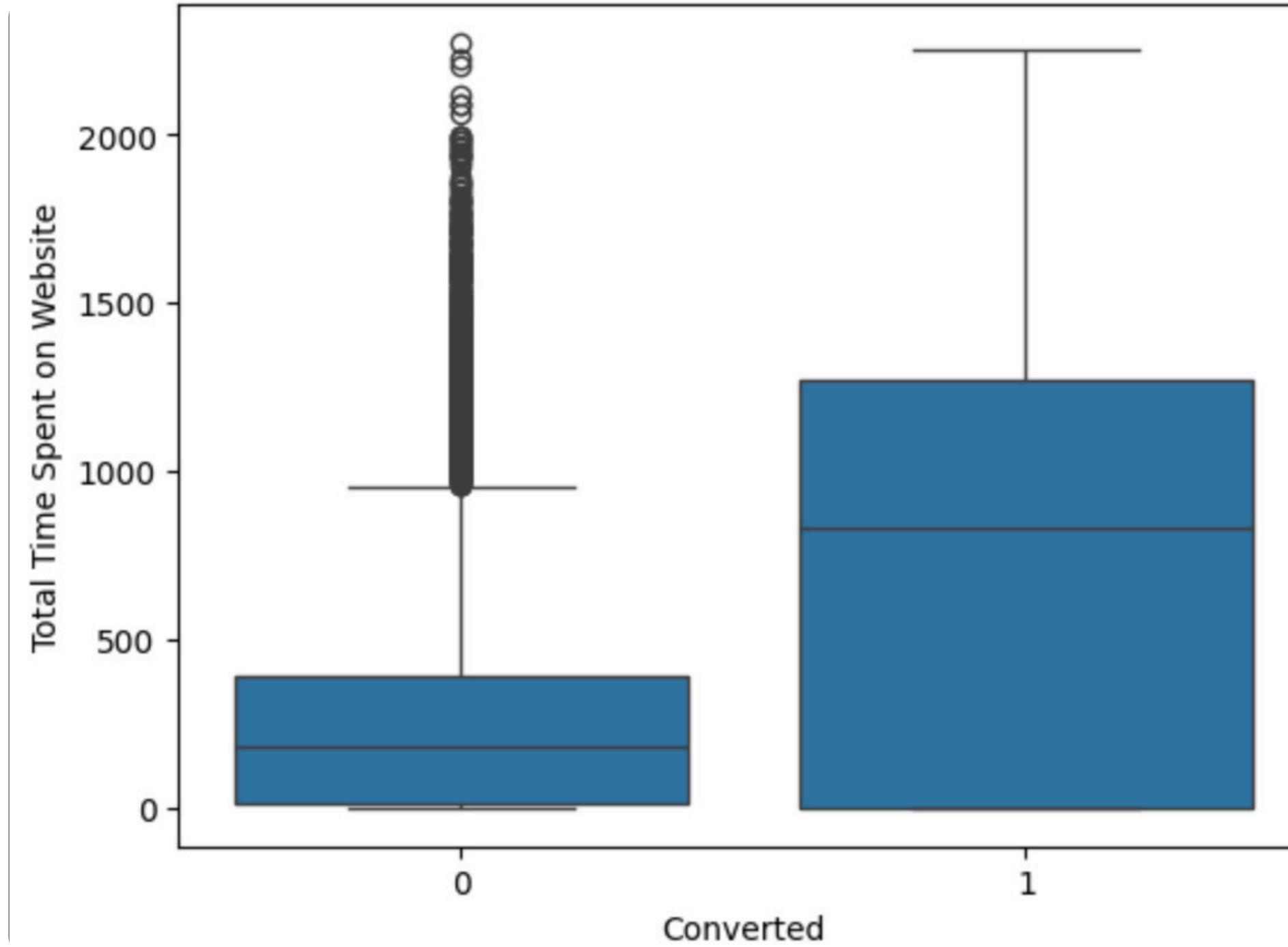
Histogram

# Histogram



- กราฟ Histogram แสดงความถี่การกระจายตัวของข้อมูล “Total Time Spent on Website” โดยแบ่งข้อมูลเป็น 20 bins ชี้ว่าช่วงให้เห็นลักษณะการกระจายตัวของข้อมูลทั้งหมด

```
# plotting boxplot w.r.t. the target variable  
sns.boxplot(y = 'Total Time Spent on Website', x = 'Converted', data = data)  
plt.show()
```



จากการ สามารถสังเกตได้ว่า ผู้ที่ใช้เวลาอยู่บนเว็บไซต์มาก มีแนวโน้มที่จะ Converted (เปลี่ยนเป็นลูกค้าจริง ๆ) มากกว่าผู้ที่ใช้เวลาน้อย



# CITY

แสดงข้อมูลเมืองที่อยู่ของกลุ่มลูกค้าที่สนใจ



# CITY

```
data['City'].value_counts()
```

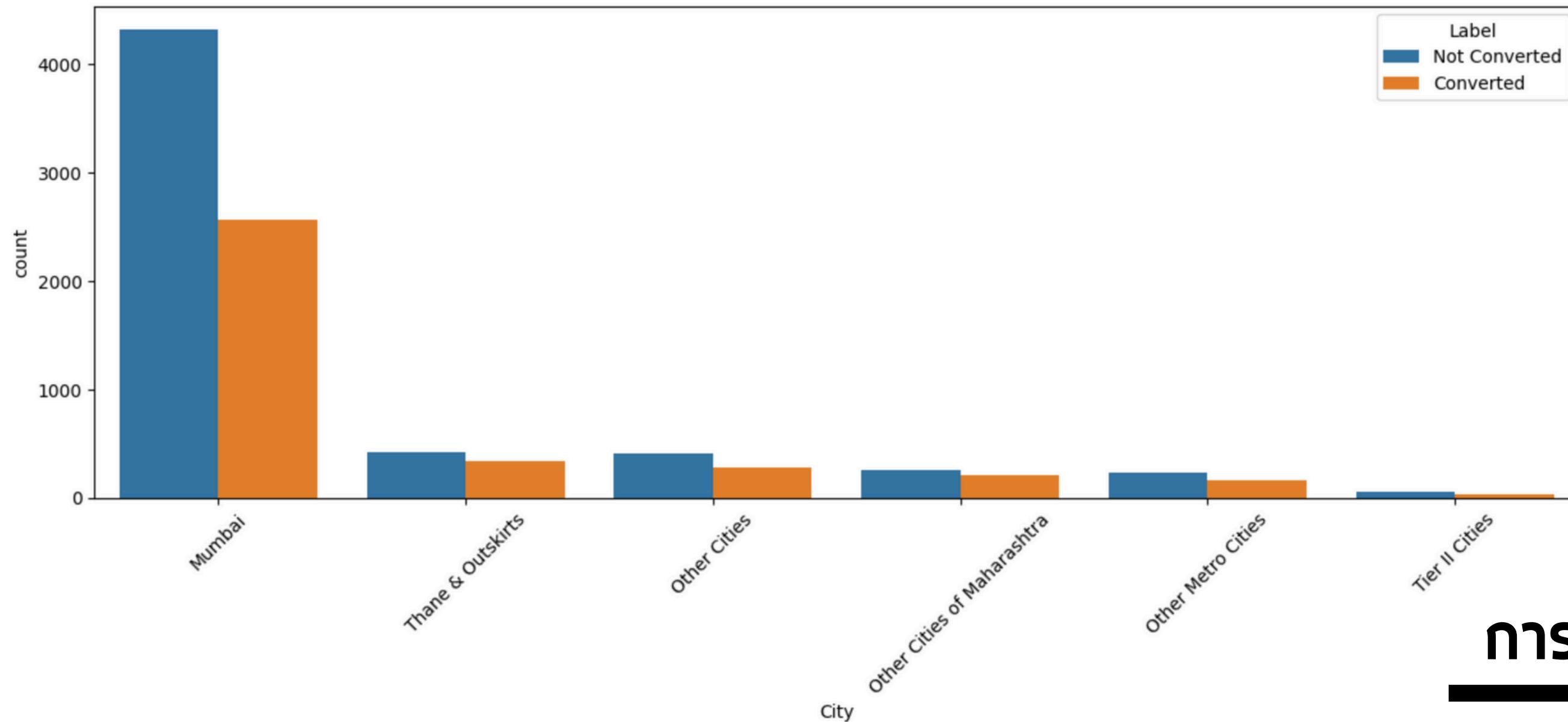
City

City	Count
Mumbai	6891
Thane & Outskirts	752
Other Cities	686
Other Cities of Maharashtra	457
Other Metro Cities	379
Tier II Cities	74

แสดงข้อมูลเมืองที่อยู่ของกลุ่มลูกค้าที่สนใจ

## สร้างกราฟ City โดยแยก Lead ที่ถูก Converted จากสี

```
# สร้างกราฟ City โดยแยก Lead ที่ถูก Converted จากสี
fig, axs = plt.subplots(figsize = (15,5))
sns.countplot(x = "City", hue = "Label", data = data, order = data['City'].value_counts().index)
xticks(rotation = 45)
plt.show()
```



## คำนวณอัตราการเปลี่ยนแปลงสถานะของลูกค้ากับเมืองที่อยู่ จากนั้นแสดงผลลอกมาเป็น DataFrame

```
● # คำนวณอัตราการเปลี่ยนแปลงสถานะของลูกค้ากับเมืองที่อยู่ จากนั้นแสดงผลลอกมาเป็น DataFrame
d = {}
for val in data['City'].unique():
    a = data[data['City'] == val]['Converted'].sum()
    b = data[data['City']==val]['Converted'].count()
    d[val] = [a, b, round(a/b*100, 2)]
pd.DataFrame.from_dict(d, orient='index').rename(columns = {0: 'Converted', 1: 'Leads',2: 'Conversion Ratio'}).sort_values(by=['Conversion Ratio'], ascending=False)
```

	Converted	Leads	Conversion Ratio
<b>Thane &amp; Outskirts</b>	338	752	44.95
<b>Other Cities of Maharashtra</b>	201	457	43.98
<b>Other Metro Cities</b>	154	379	40.63
<b>Other Cities</b>	276	686	40.23
<b>Mumbai</b>	2566	6891	37.24
<b>Tier II Cities</b>	25	74	33.78

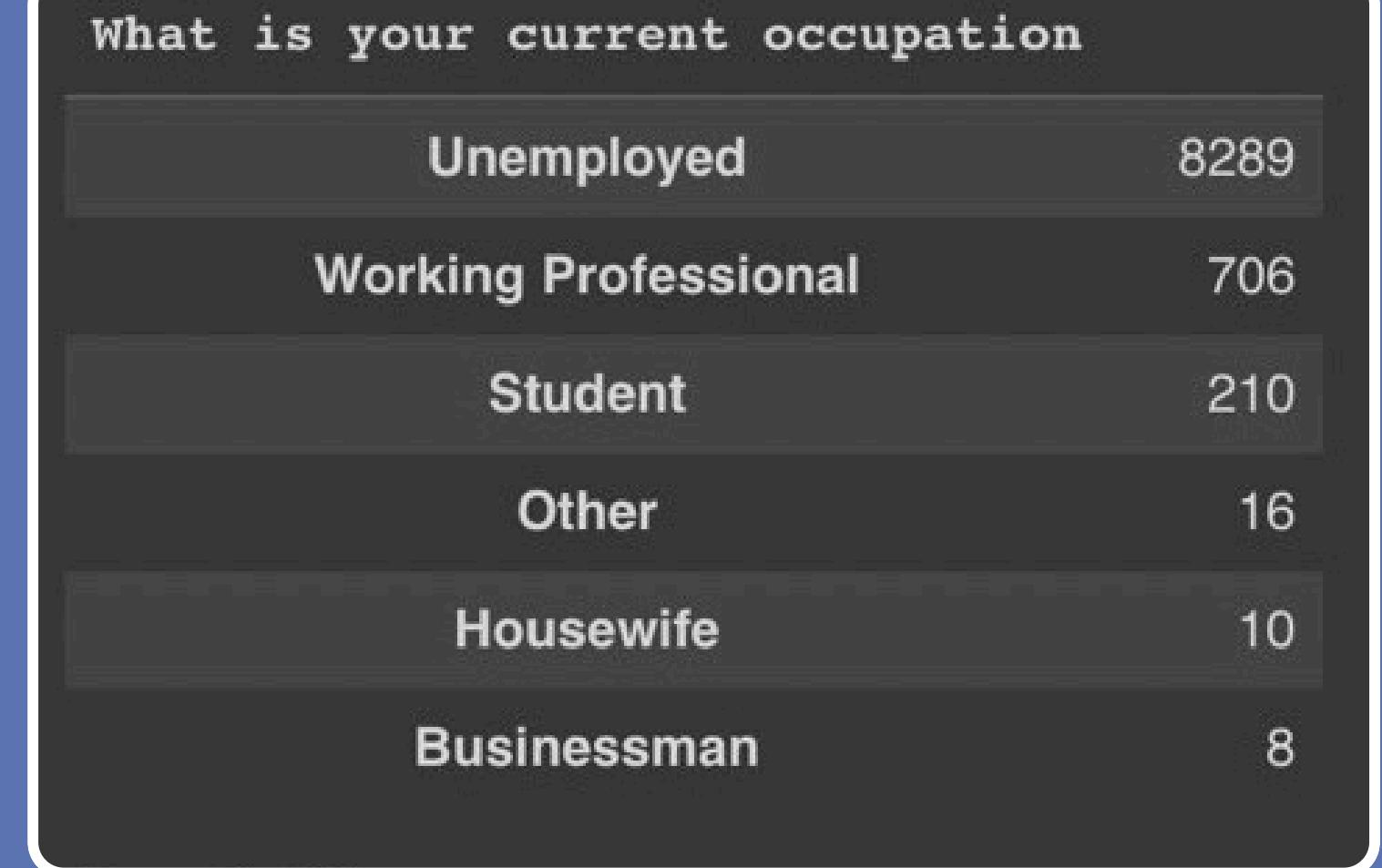


# What is your current occupation

```
data['What is your current occupation'].value_counts()
```



แสดงข้อมูลอาชีพที่กำroyในปัจจุบัน



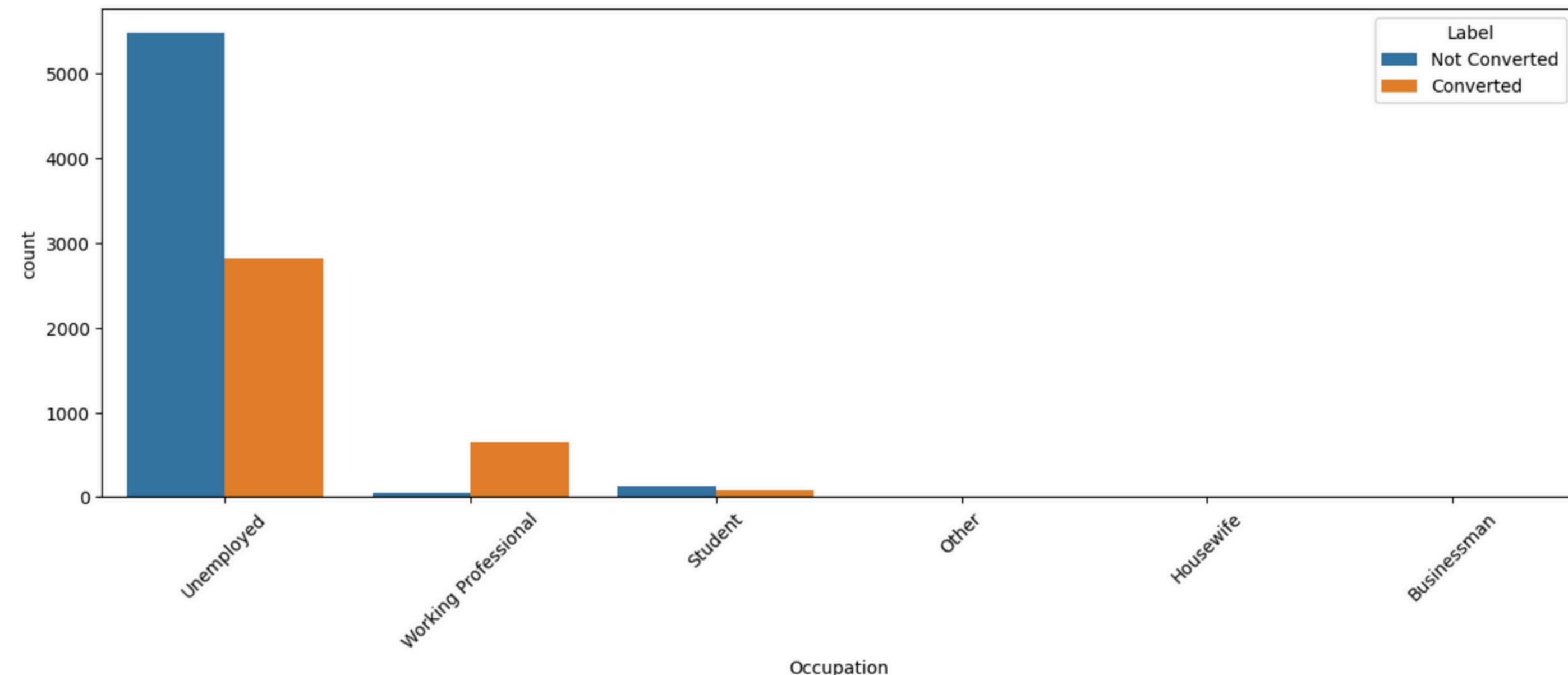
# Current Occupation

GROUP 20

# สร้างกราฟ Occupation โดยแยก Lead ที่ถูก Converted จากสี

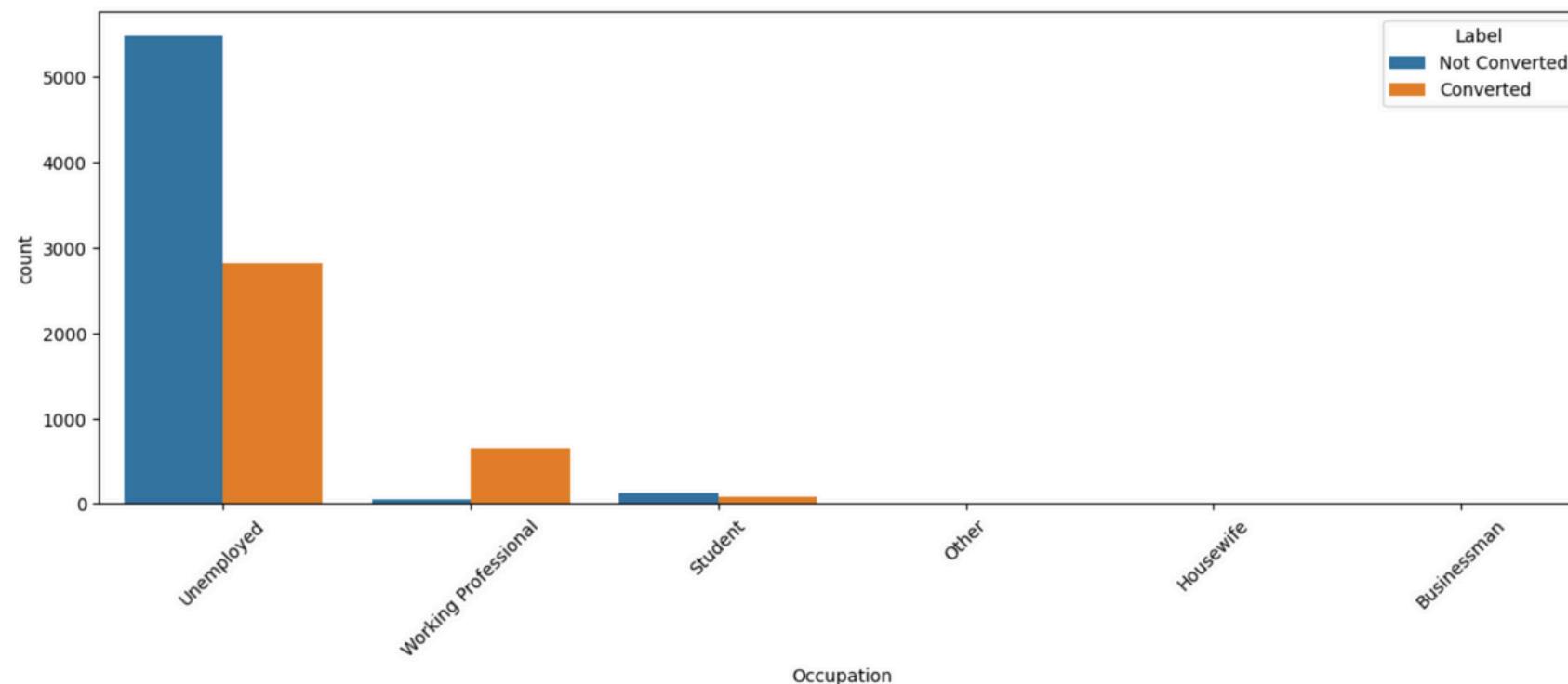
# เปลี่ยนชื่อจาก What is your current occupation เป็น Occupation

```
▶ # สร้างกราฟ Occupation โดยแยก Lead ที่ถูก Converted จากสี
fig, axs = plt.subplots(figsize = (15,5))
sns.countplot(x = "What is your current occupation", hue="Label", data=data['What is your current occupation'].value_counts().index)
plt.xlabel('Occupation') #เปลี่ยนชื่อจาก What is your current occupation เป็น Occupation
xticks(rotation = 45)
plt.show()
```



# Current Occupation

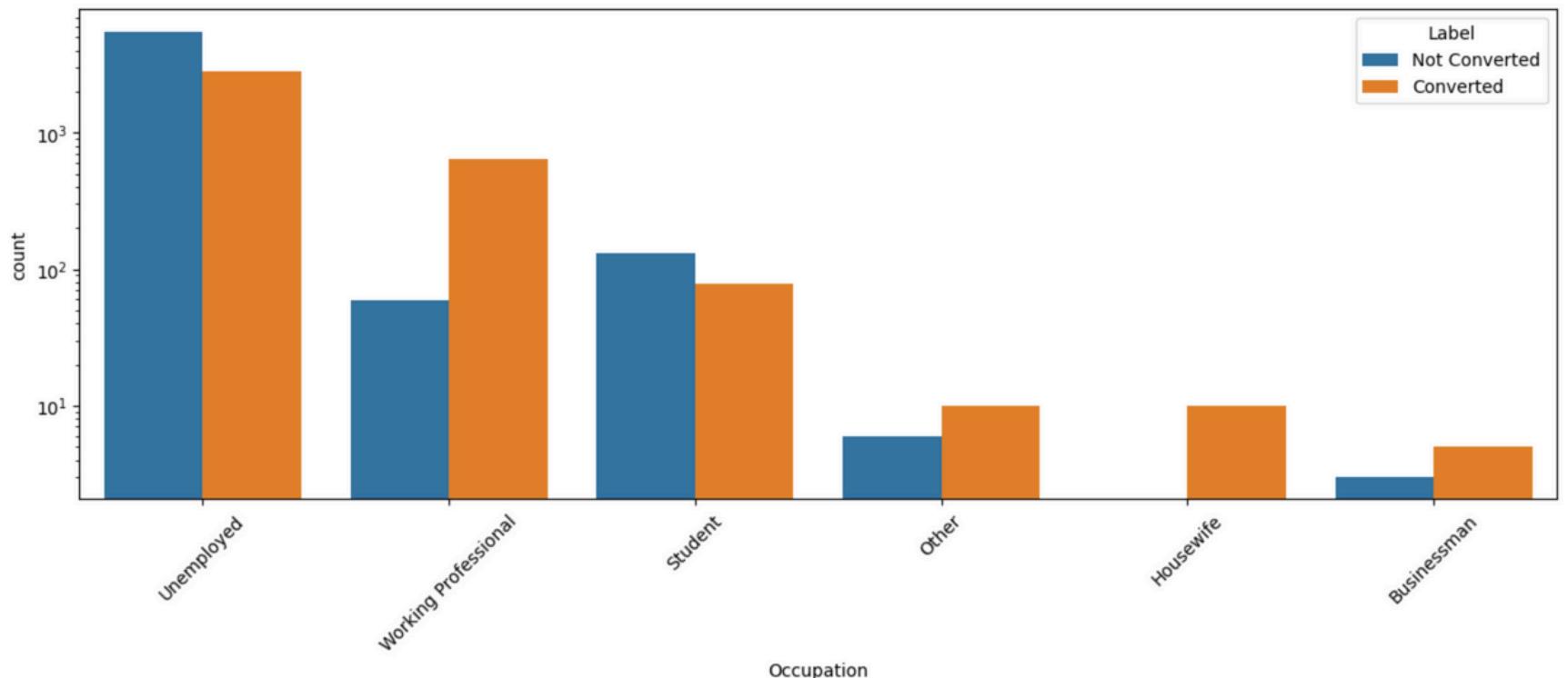
GROUP 20



จากการฟกร่อนหน้าจะเห็นได้ว่าชุดข้อมูลมีความถ่วงกันอย่างมากทำให้มองเห็นได้ยาก  
ซึ่งสามารถแก้ไขได้โดยใช้ log Scale มาช่วยให้มองสะพานมากขึ้นได้

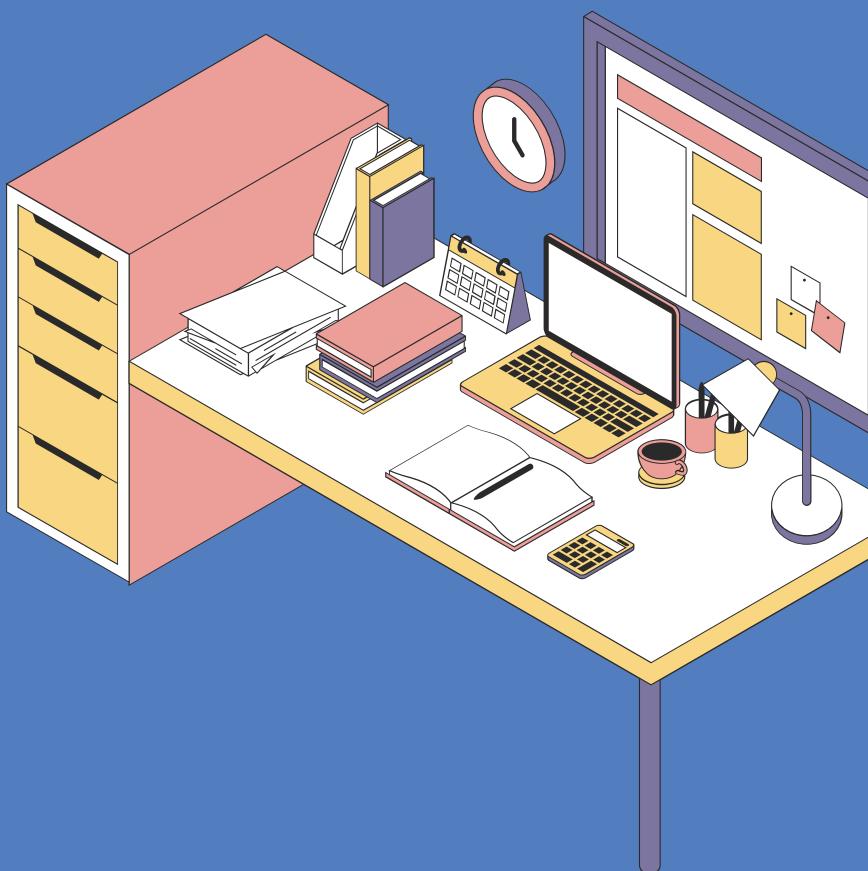
## Plt.yscale('log')

# ใช้ logarithmic scale บนแกน y



```
▶ fig, axs = plt.subplots(figsize=(15, 5))
sns.countplot(x="What is your current occupation", hue="Label", data=data, order=data['What is your current occupation'].value_counts().index)
plt.yscale('log') # ใช้ logarithmic scale บนแกน y
plt.xlabel('Occupation') # เปลี่ยนชื่อจาก What is your current occupation เป็น Occupation
plt.xticks(rotation=45)
plt.show()
```

# Current Occupation



คำนวณอัตราการเปลี่ยนแปลงสถานะของลูกค้ากับอาชีพปัจจุบันของลูกค้า

```
# คำนวณอัตราการเปลี่ยนแปลงสถานะของลูกค้ากับอาชีพปัจจุบันของลูกค้า จากนั้นแสดงผลออกมายัง DataFrame
d = {}
for val in data['What is your current occupation'].unique():
    a = data[data['What is your current occupation'] == val]['Converted'].sum()
    b = data[data['What is your current occupation'] == val]['Converted'].count()
    d[val] = [a, b, round(a/b*100, 2)]
pd.DataFrame.from_dict(d, orient='index').rename(columns = {0: 'Converted', 1: 'Leads', 2: 'Conversion Ratio'}).sort_values(by=['Conversion Ratio'], ascending=False)
```

จากนั้นแสดงผลออกมายัง DataFrame

	Converted	Leads	Conversion Ratio
<b>Housewife</b>	10	10	100.00
<b>Working Professional</b>	647	706	91.64
<b>Businessman</b>	5	8	62.50
<b>Other</b>	10	16	62.50
<b>Student</b>	78	210	37.14
<b>Unemployed</b>	2810	8289	33.90

# Specialization



Specialization	
Finance Management	4356
Human Resource Management	848
Marketing Management	838
Operations Management	503
Business Administration	403
IT Projects Management	366
Supply Chain Management	349
Banking, Investment And Insurance	338
Media and Advertising	203
Travel and Tourism	203
International Business	178
Healthcare Management	159
Hospitality Management	114
E-COMMERCE	111
Retail Management	100
Rural and Agribusiness	73
E-Business	57
Services Excellence	40

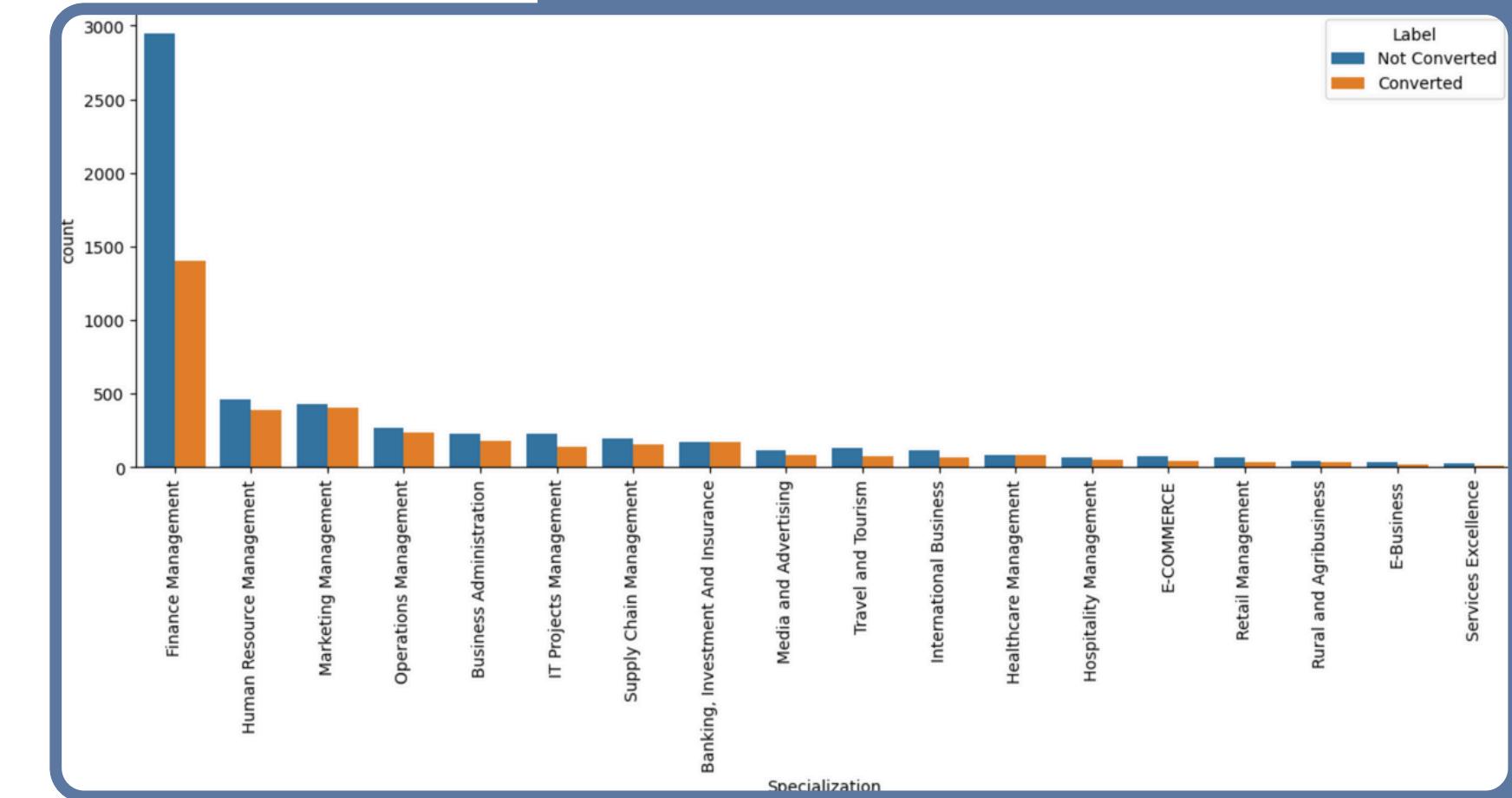
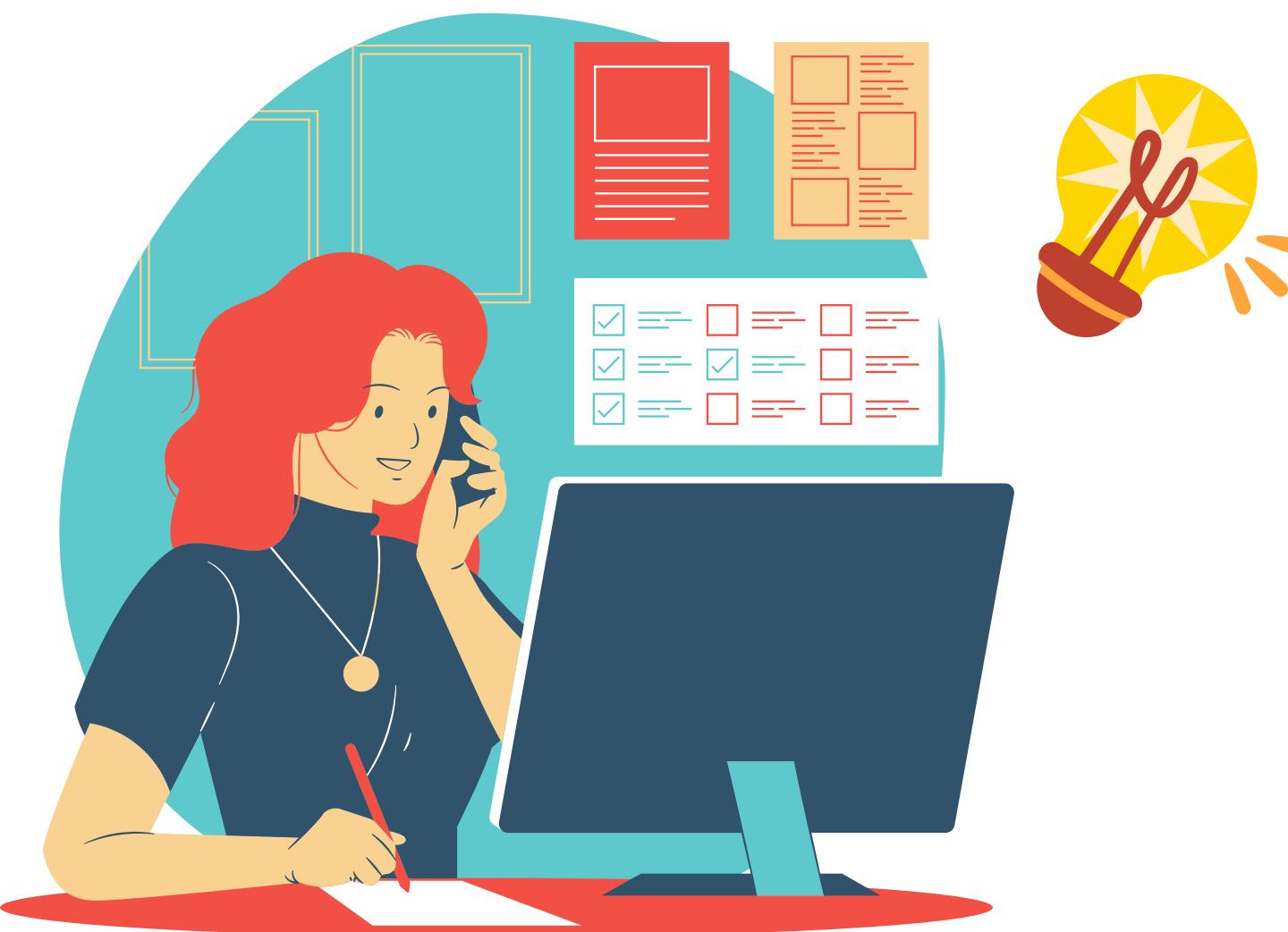
dtype: int64



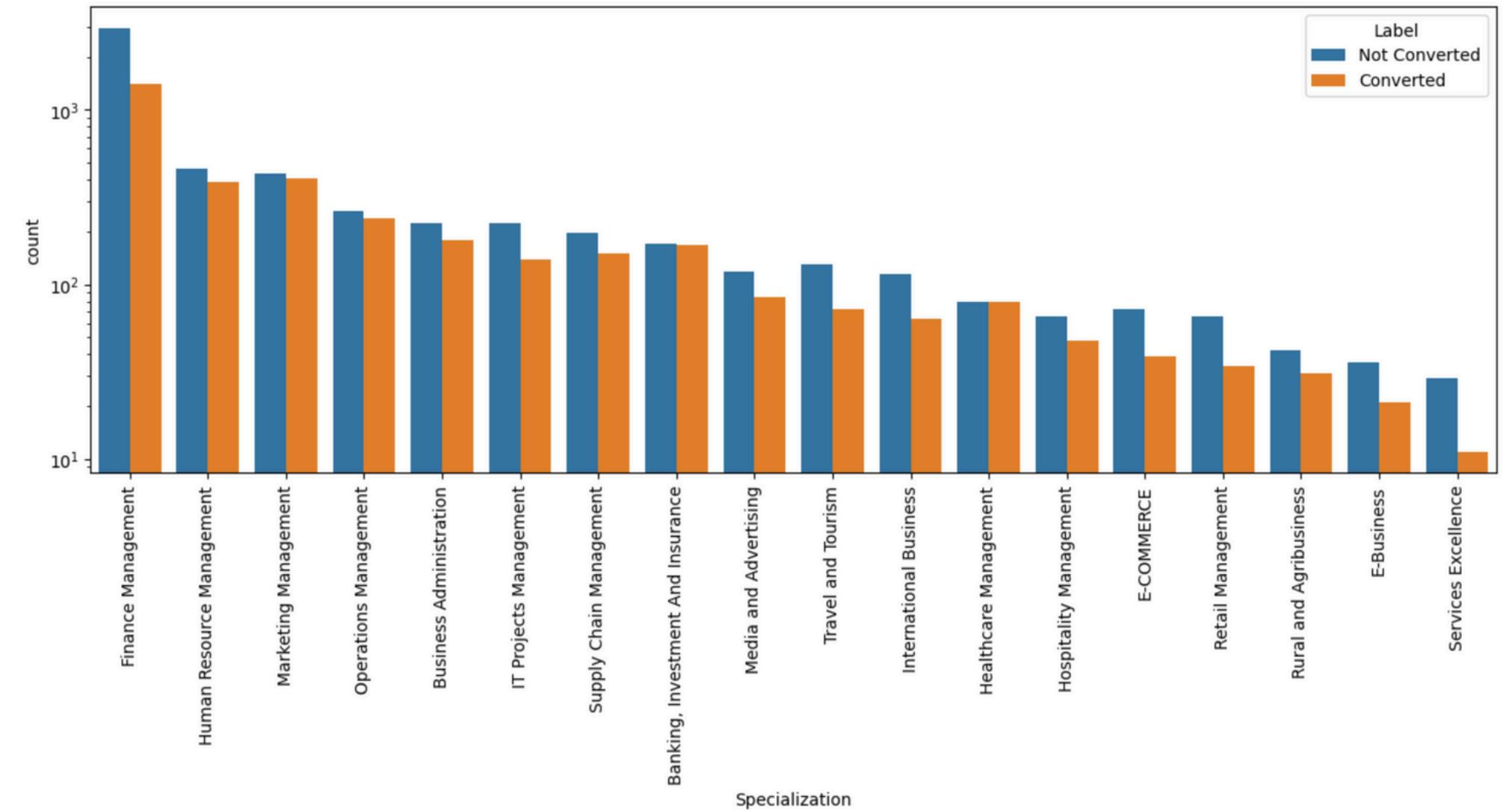
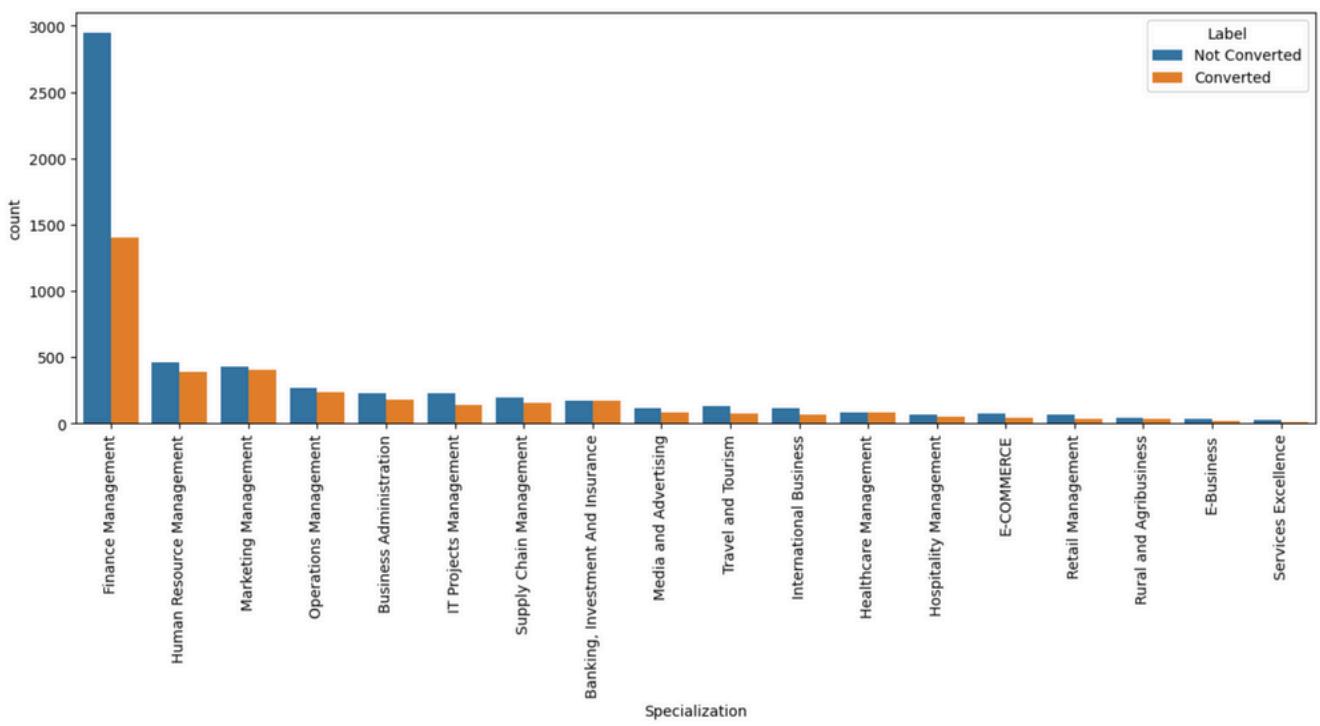
data['Specialization'].value\_counts()

# Specialization

```
▶ # สร้างกราฟ Specialization โดยแยก Lead ที่ถูก Converted จากสี  
fig, axs = plt.subplots(figsize = (15,5))  
  
sns.countplot(x = "Specialization", hue = "Label", data = data, order = data['Specialization'].value_counts().index)  
  
xticks(rotation = 90)  
  
plt.show()
```



# Specialization



Plt.yscale('log')

# Specialization

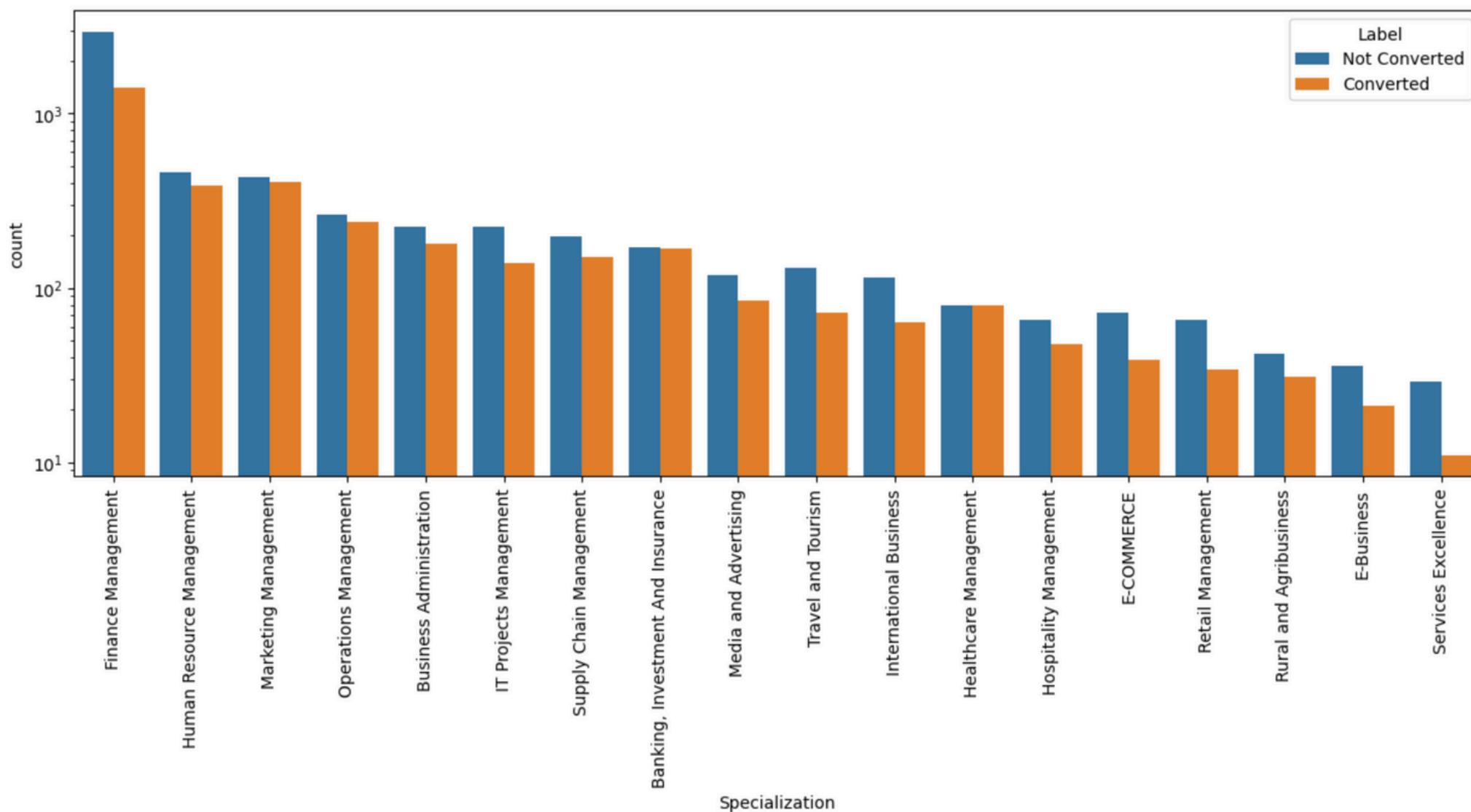
▶ #คำนวณอัตราการเปลี่ยนแปลงสถานะของลูกค้า และแสดงผลออกมารูป DataFrame

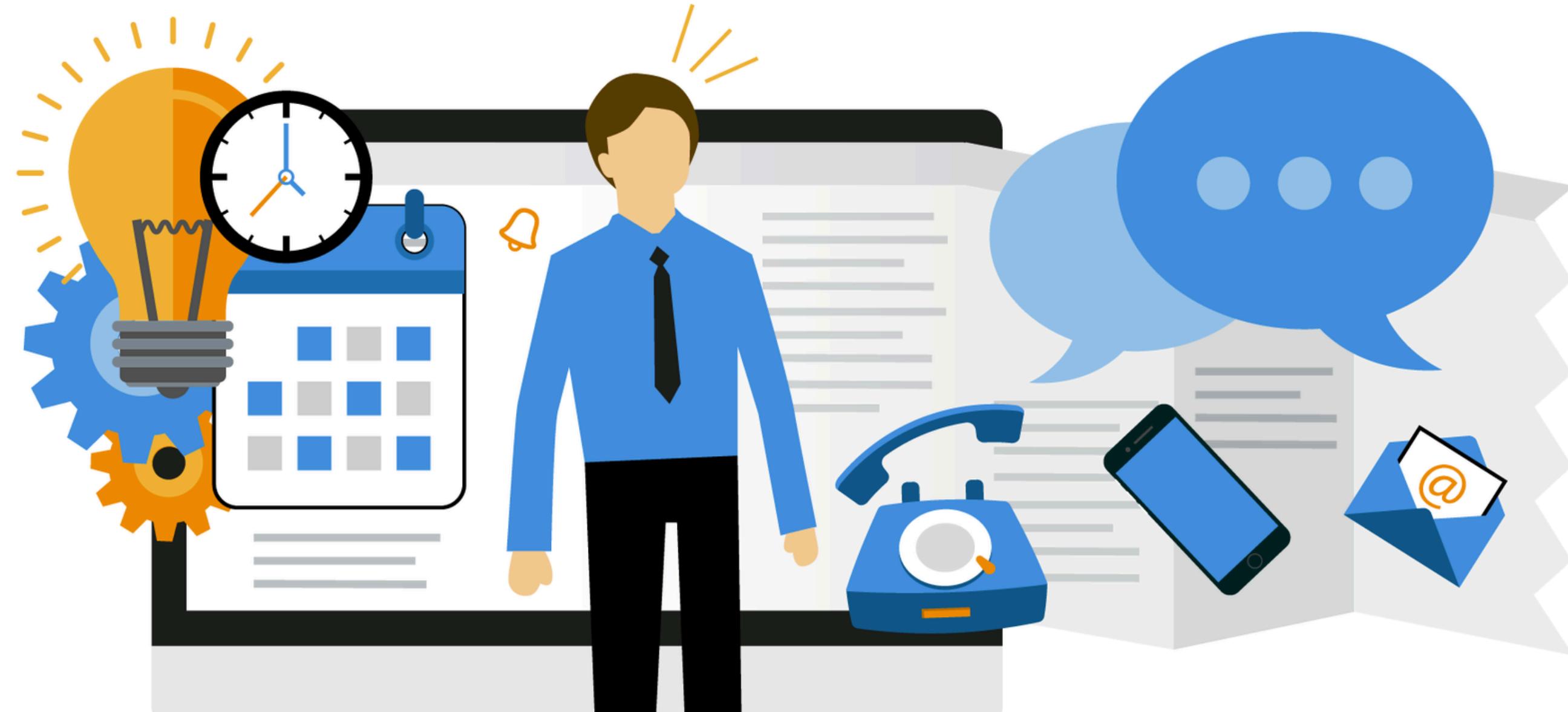
```

d = {}
for val in data['Specialization'].unique():
    a = data[data['Specialization'] == val]['Converted'].sum()
    b = data[data['Specialization']==val]['Converted'].count()
    d[val] = [a, b, round(a/b*100, 2)]
pd.DataFrame.from_dict(d, orient='index').rename(columns = {0: 'Converted', 1: 'Leads',2: 'Conversion Ratio'}).sort_values(by=['Conversion Ratio'], ascending=False)

```

	Converted	Leads	Conversion Ratio
Healthcare Management	79	159	49.69
Banking, Investment And Insurance	167	338	49.41
Marketing Management	408	838	48.69
Operations Management	238	503	47.32
Human Resource Management	388	848	45.75
Business Administration	179	403	44.42
Supply Chain Management	151	349	43.27
Rural and Agribusiness	31	73	42.47
Hospitality Management	48	114	42.11
Media and Advertising	85	203	41.87
IT Projects Management	140	366	38.25
E-Business	21	57	36.84
International Business	64	178	35.96
Travel and Tourism	72	203	35.47
E-COMMERCE	39	111	35.14
Retail Management	34	100	34.00
Finance Management	1405	4356	32.25
Services Excellence	11	40	27.50





# Last Activity

# Last Activity

```
data['Last Activity'].value_counts()
```

แสดงข้อมูลกิจกรรมล่าสุดก่อนเข้าชมเว็บไซด์

Last Activity	
Email Opened	3540
SMS Sent	2745
Olarck Chat Conversation	973
Page Visited on Website	640
Converted to Lead	428
Email Bounced	325
Email Link Clicked	267
Form Submitted on Website	116

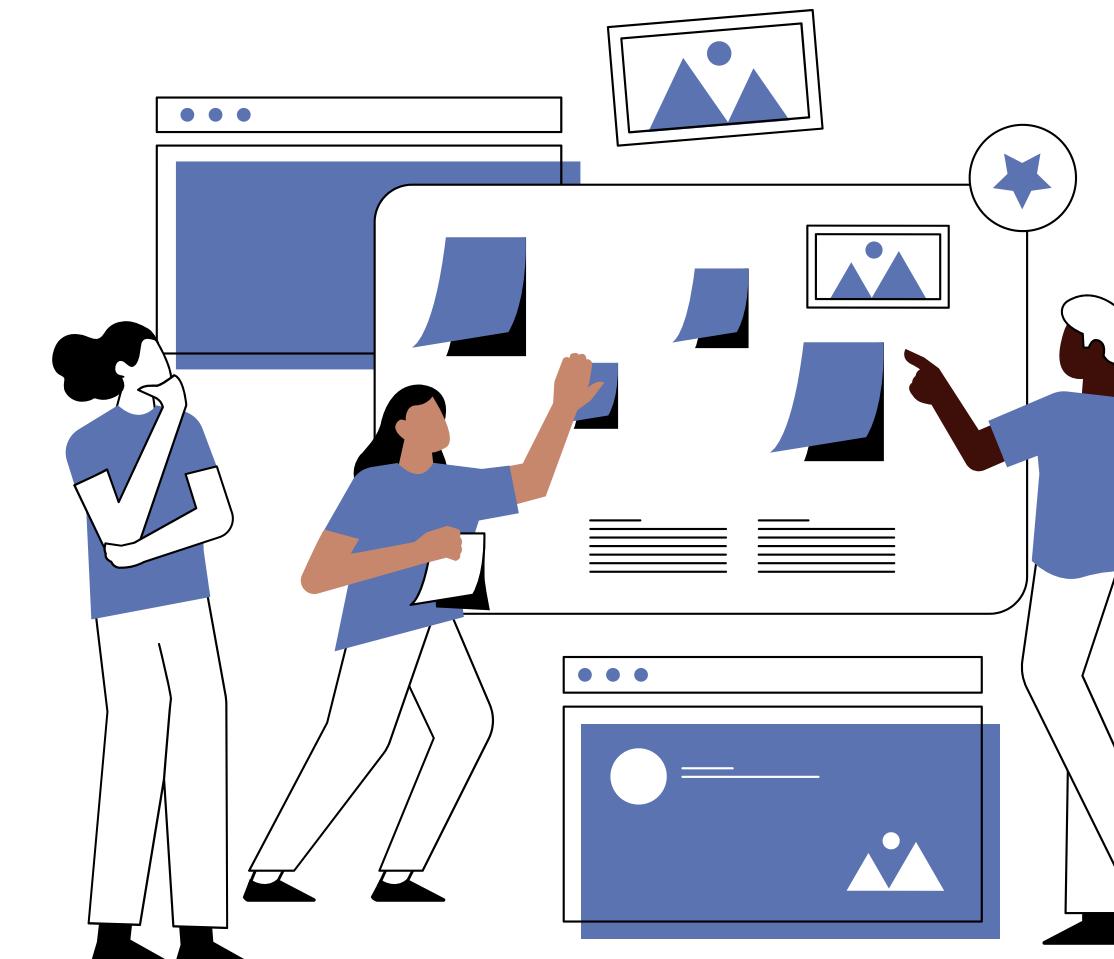
Unreachable	93
Unsubscribed	61
Had a Phone Conversation	30
Approached upfront	9
View in browser link Clicked	6
Email Received	2
Email Marked Spam	2
Visited Booth in Tradeshow	1
Resubscribed to emails	1



# Last Activity

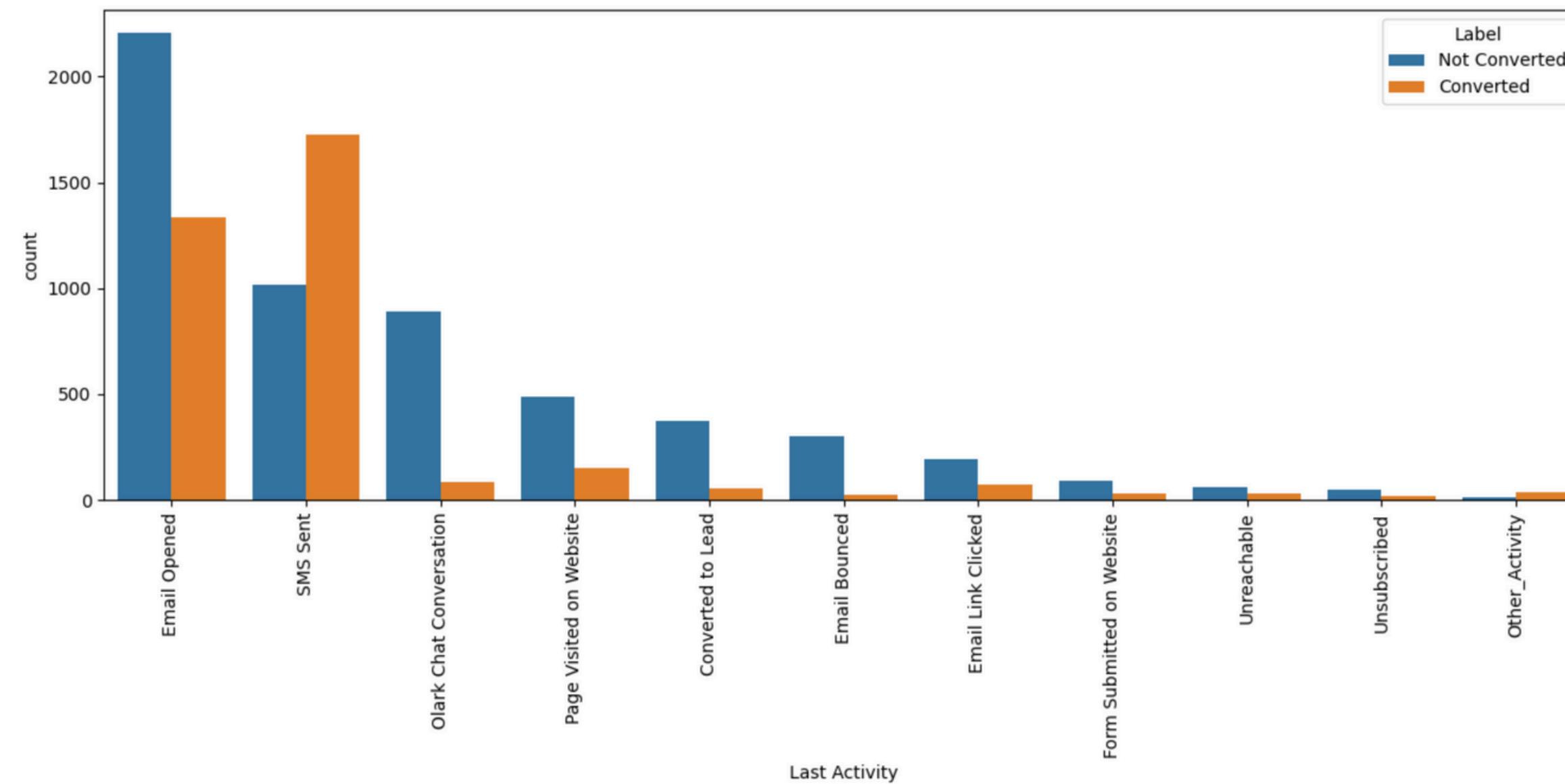
<b>Unreachable</b>	93
<b>Unsubscribed</b>	61
<b>Had a Phone Conversation</b>	30
<b>Approached upfront</b>	9
<b>View in browser link Clicked</b>	6
<b>Email Received</b>	2
<b>Email Marked Spam</b>	2
<b>Visited Booth in Tradeshow</b>	1
<b>Resubscribed to emails</b>	1

```
# เปลี่ยนชุดข้อมูลที่มีน้อยกว่า 30 เป็น Other_Activity  
data['Last Activity'] = data['Last Activity'].replace(['Had a Phone Conversation',  
'View in browser link Clicked',  
'Visited Booth in Tradeshow',  
'Approached upfront',  
'Resubscribed to emails',  
'Email Received',  
'Email Marked Spam'], 'Other_Activity')
```

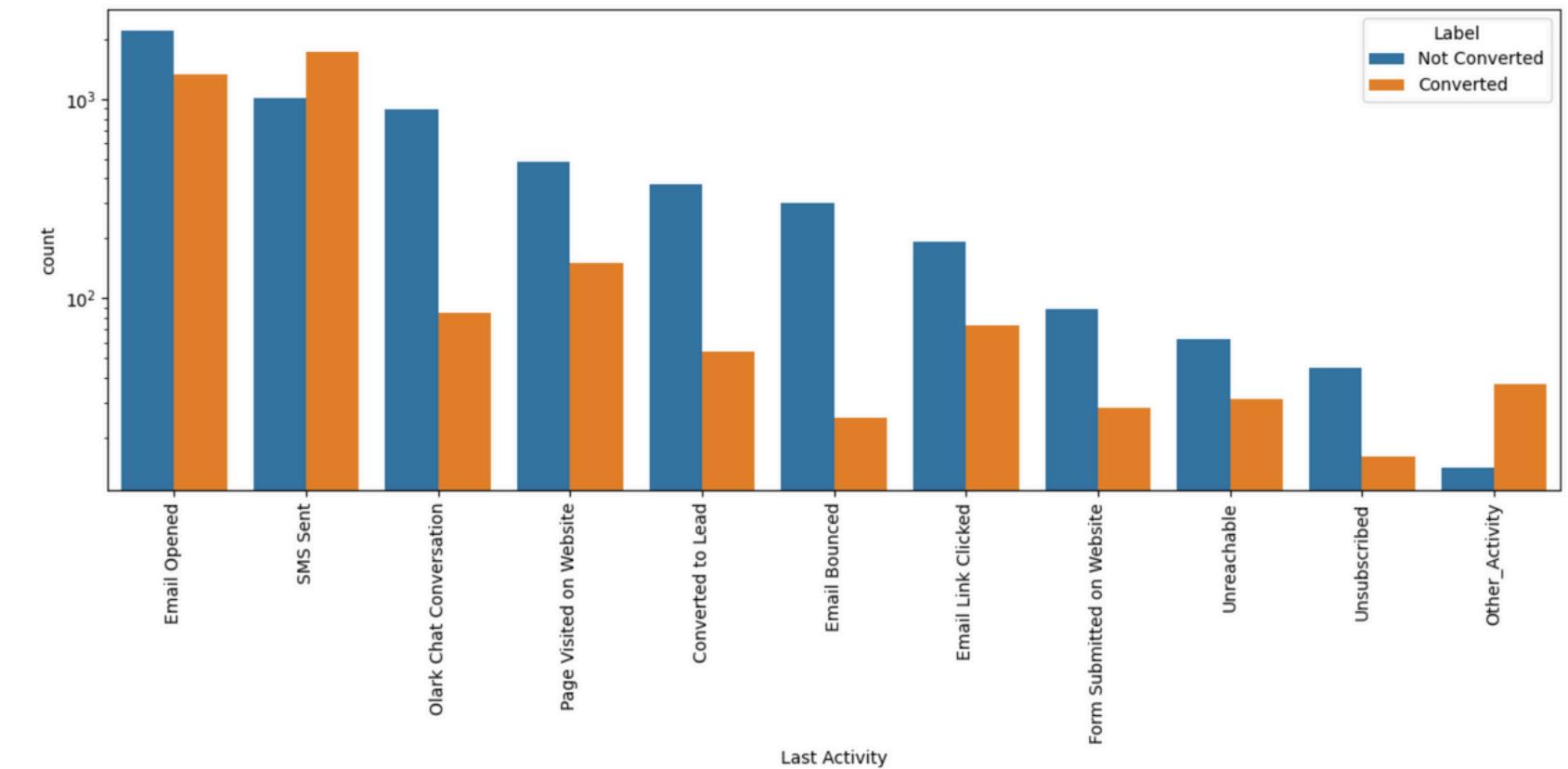
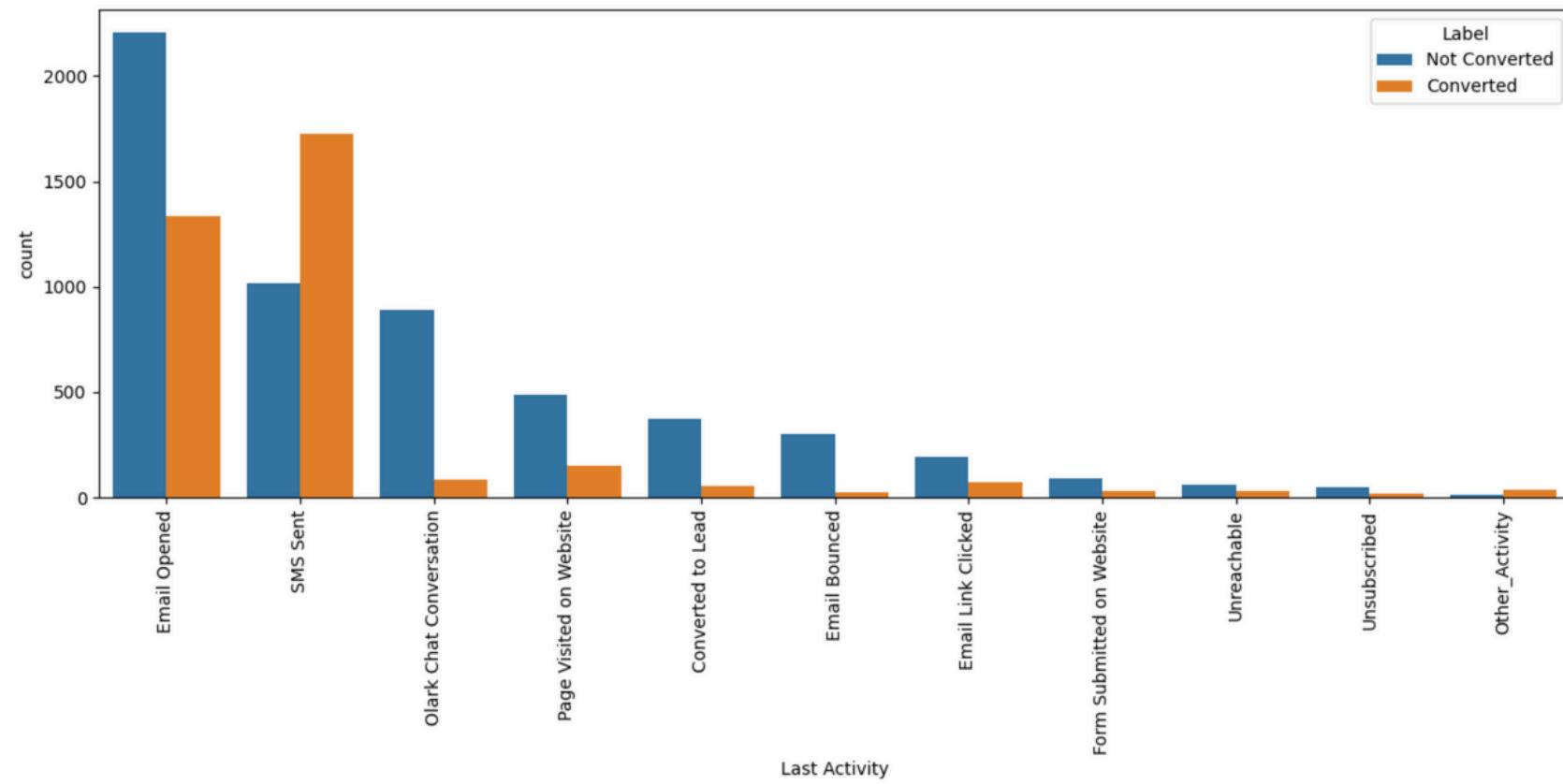


# Last Activity

```
▶ # สร้างกราฟ Last Activity โดยแยก Lead ที่ถูก Converted จากสี
    fig, axs = plt.subplots(figsize=(15,5))
    sns.countplot(x="Last Activity", hue="Label", data=data, order=data['Last Activity'].value_counts().index)
```



# ข้อดีและข้อเสียการใช้ log scale



→  
Plt.yscale('log')

# Last Activity

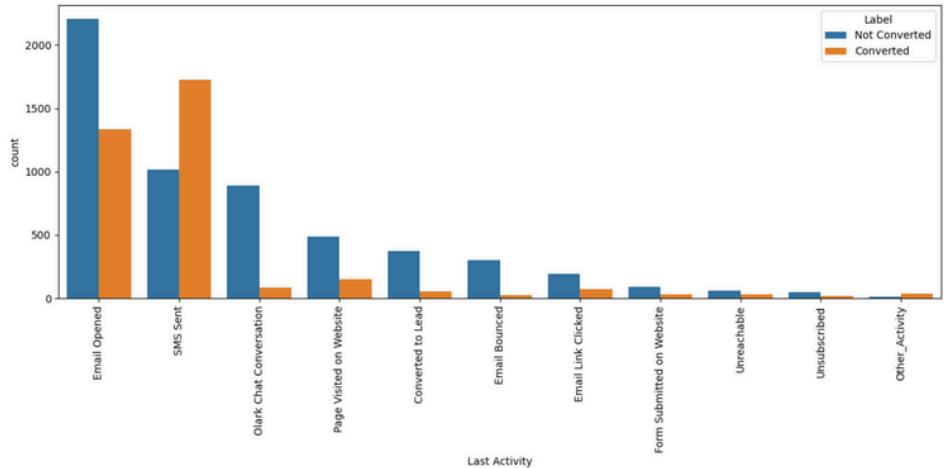
```
▶ # กรองข้อมูลเพื่อนำไปสร้างเป็นกราฟ
last_activities = data['Last Activity'].value_counts().index[-5:] #เลือก 5 คอลัมน์ท้าย
filtered_data = data[data['Last Activity'].isin(last_activities)] #กรอง DataFrame เพื่อรวมเฉพาะกิจกรรมที่เลือก
remaining_data = data[~data['Last Activity'].isin(last_activities)] #กำหนดค่าคอลัมน์ที่ไม่ได้เลือกไว้เพื่อนำไปแสดงในกราฟแรก

# สร้างกราฟ Last Activity โดยกราฟแรกจะแสดง 6 คอลัมน์แรก
fig, axs = plt.subplots(1,2, figsize = (20,6.5)) #
sns.countplot(x="Last Activity", hue="Label", data=remaining_data,
               order=remaining_data['Last Activity'].value_counts().index, ax=axs[0])
axs[0].tick_params(axis='x', rotation=45)

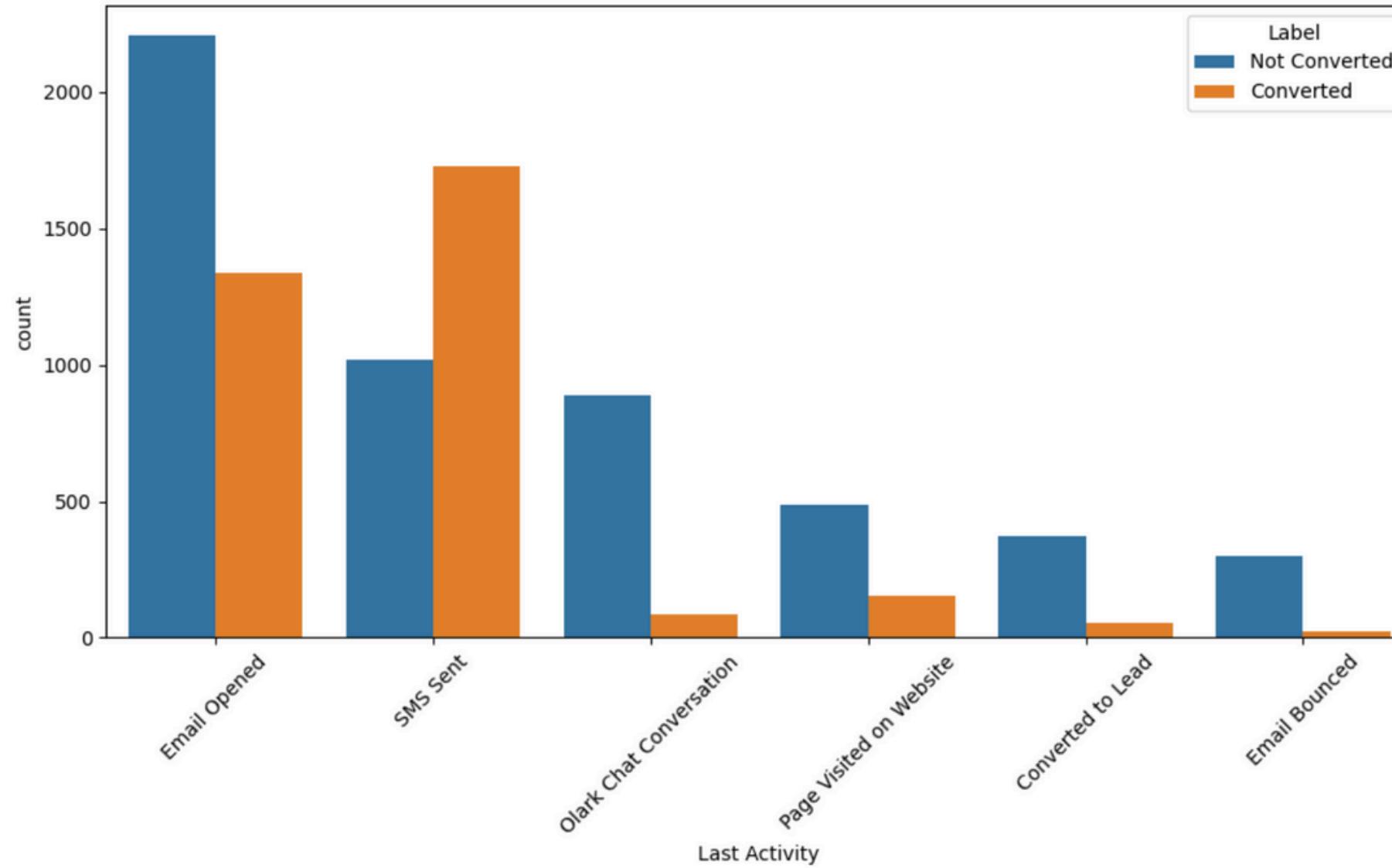
# สร้างกราฟ Last Activity โดยกราฟแรกจะแสดง 5 คอลัมน์สุดท้าย
sns.countplot(x="Last Activity", hue="Label", data=filtered_data, order=last_activities, ax = axs[1])
plt.xticks(rotation=45)

plt.tight_layout() #จัด Layout ให้เท่ากัน
plt.show()
```

# กราฟเดิม

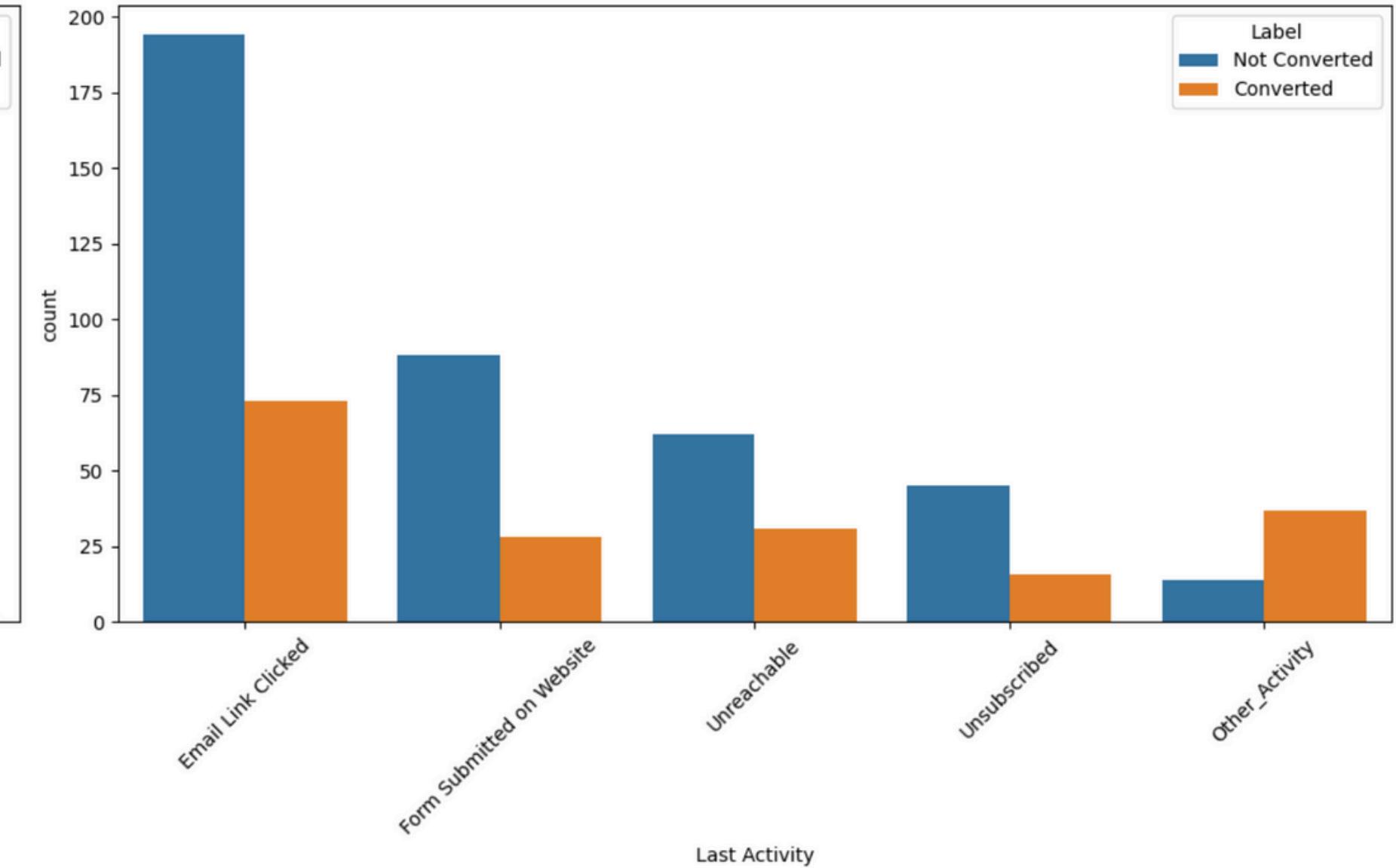


Scale 0-2000



6 ชุดข้อมูลแรกที่มีค่ามาก

Scale 0-200



5 ชุดข้อมูลแรกที่มีค่ามาก

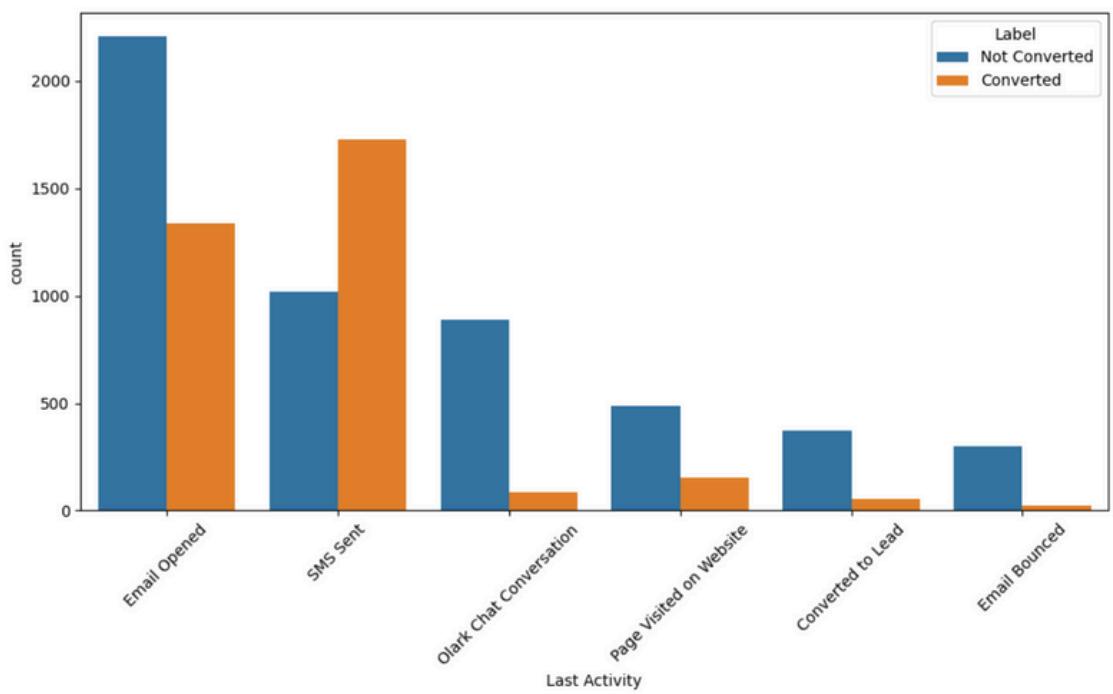
# Last Activity

GROUP 20

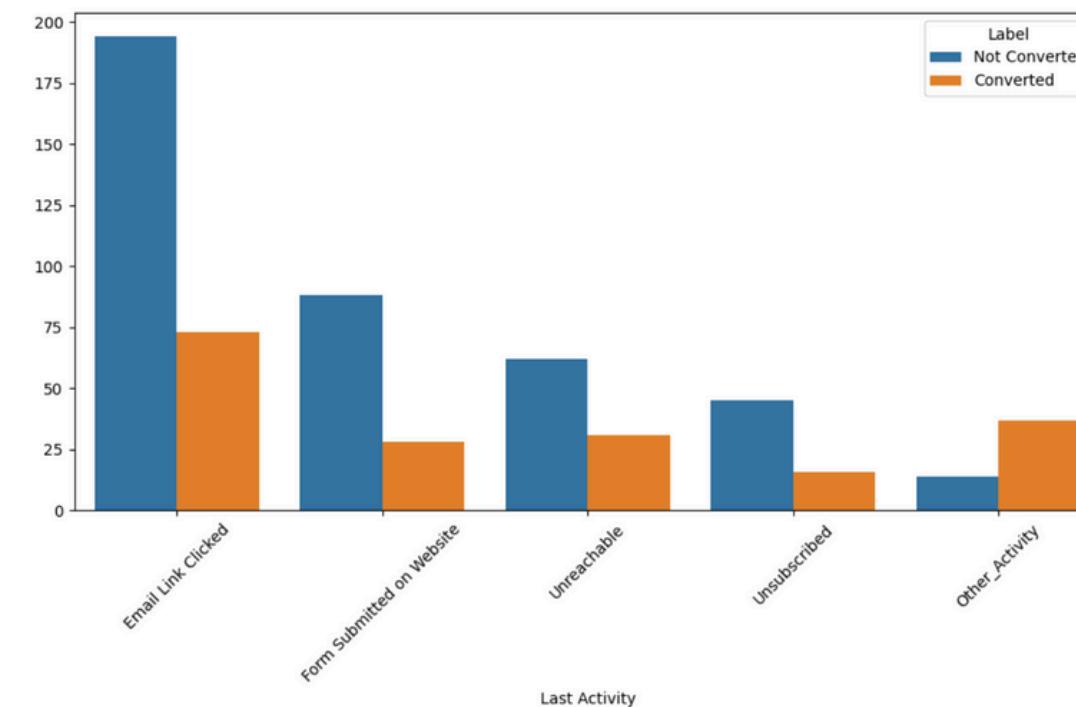
▶ # คำนวณอัตราการเปลี่ยนแปลงสถานะของลูกค้าและแสดงผลออกมาเป็น DataFrame

```
d = {}
for val in data['Last Activity'].unique():
    a = data[data['Last Activity'] == val]['Converted'].sum()
    b = data[data['Last Activity'] == val]['Converted'].count()
    d[val] = [a, b, round(a/b*100, 2)]
pd.DataFrame.from_dict(d, orient='index').rename(columns = {0: 'Converted', 1: 'Leads',2: 'Conversion Ratio'}).sort_values(by=['Conversion Ratio'], ascending=False)
```

Scale 0-2000



Scale 0-200



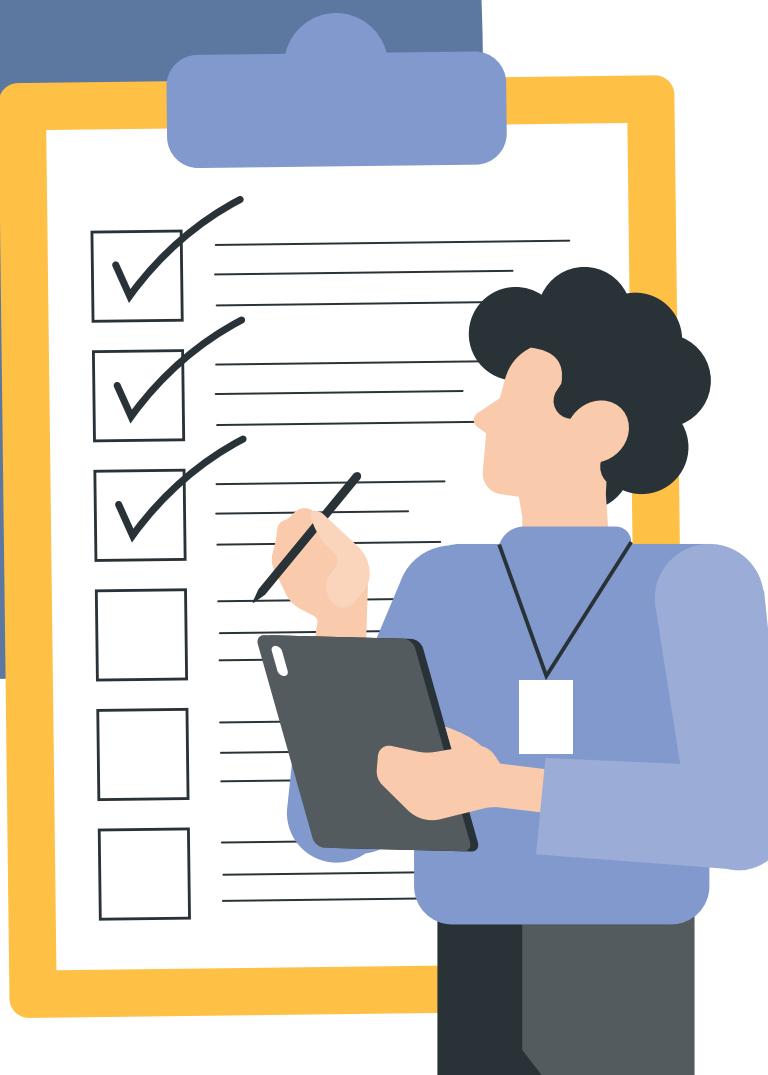
Converted Leads Conversion Ratio

Other_Activity	Converted	Leads	Conversion Ratio
Other_Activity	37	51	72.55
SMS Sent	1727	2745	62.91
Email Opened	1334	3540	37.68
Unreachable	31	93	33.33
Email Link Clicked	73	267	27.34
Unsubscribed	16	61	26.23
Form Submitted on Website	28	116	24.14
Page Visited on Website	151	640	23.59
Converted to Lead	54	428	12.62
Olark Chat Conversation	84	973	8.63
Email Bounced	25	325	7.69

# Exploratory Data Analysis (EDA) on Lead Scoring Dataset

---  
GROUP 20

EDA on Lead Scoring Dataset



# ★ REFERENCES



<https://www.analyticsvidhya.com/blog/2022/09/exploratory-data-analysis-eda-on-lead-scoring-dataset/>

<https://blog.skooldio.com/increase-your-customer-with-exploratory-data-analysis>



# รายชื่อสมาชิก

65050003	กชพร ตี๊ะเหมียว
65050216	ໂছດິກະ ດັ່ງພົມເຈົ້າ
65050241	ນກັກສ ສີຣິນາ
65050470	ນາຮີຮັດນິ ພລວມ
65050479	ນິສິກູນິ ກູ່ງາມຂຶ້ນ
65050630	ພິດພັສສາ ທາຍາດ
65050671	ກັດທິຣາ ໃຈດີ
65050678	ກັກຮວດ ສຣີສໍາຮາລູບໍາຮຸງ
65050790	ວຽນຖຸດີ ຜລາຮັກເງົ

THANK YOU!!