In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

path=r'C:\Users\DELL\Documents\project\data set\train_LZdllcl.csv'
train_df=pd.read_csv(path)
train_df
```

Out[1]:

| | employee_id | department | region | education | gender | recruitment_channel | n |
|---|---|---|---|---|---|---|---|
| **0** | 65438 | Sales & Marketing | region_7 | Master's & above | f | sourcing | |
| **1** | 65141 | Operations | region_22 | Bachelor's | m | other | |
| **2** | 7513 | Sales & Marketing | region_19 | Bachelor's | m | sourcing | |
| **3** | 2542 | Sales & Marketing | region_23 | Bachelor's | m | other | |
| **4** | 48945 | Technology | region_26 | Bachelor's | m | other | |
| **...** | ... | ... | ... | ... | ... | ... | |
| **54803** | 3030 | Technology | region_14 | Bachelor's | m | sourcing | |
| **54804** | 74592 | Operations | region_27 | Master's & above | f | other | |
| **54805** | 13918 | Analytics | region_1 | Bachelor's | m | other | |
| **54806** | 13614 | Sales & Marketing | region_9 | NaN | m | sourcing | |
| **54807** | 51526 | HR | region_22 | Bachelor's | m | other | |

54808 rows × 14 columns

In [2]:
```
cat_columns=train_df.select_dtypes(include='object').columns
num_columns=train_df.select_dtypes(exclude='object').columns
```

In [3]:
```
cat_columns
```

Out[3]:
```
Index(['department', 'region', 'education', 'gender', 'recruitment_channel'], d
type='object')
```

In [4]:
```
num_columns
```

Out[4]:
```
Index(['employee_id', 'no_of_trainings', 'age', 'previous_year_rating',
       'length_of_service', 'KPIs_met >80%', 'awards_won?',
       'avg_training_score', 'is_promoted'],
      dtype='object')
```

In [5]:
```
train_df[['department']]
```

Out[5]:

| | department |
|---|---|
| **0** | Sales & Marketing |
| **1** | Operations |
| **2** | Sales & Marketing |
| **3** | Sales & Marketing |
| **4** | Technology |
| **...** | ... |
| **54803** | Technology |
| **54804** | Operations |
| **54805** | Analytics |
| **54806** | Sales & Marketing |
| **54807** | HR |

54808 rows × 1 columns

In [6]:
```
print(train_df['department'].unique())
print(len(train_df['department'].unique()))
```

```
['Sales & Marketing' 'Operations' 'Technology' 'Analytics' 'R&D'
 'Procurement' 'Finance' 'HR' 'Legal']
9
```

In [7]:
```
cdf=train_df['department'].value_counts()
cdf
```

Out[7]:
```
department
Sales & Marketing    16840
Operations           11348
Technology            7138
Procurement           7138
Analytics             5352
Finance               2536
HR                    2418
Legal                 1039
R&D                    999
Name: count, dtype: int64
```

In [8]:
```
type(cdf)
```

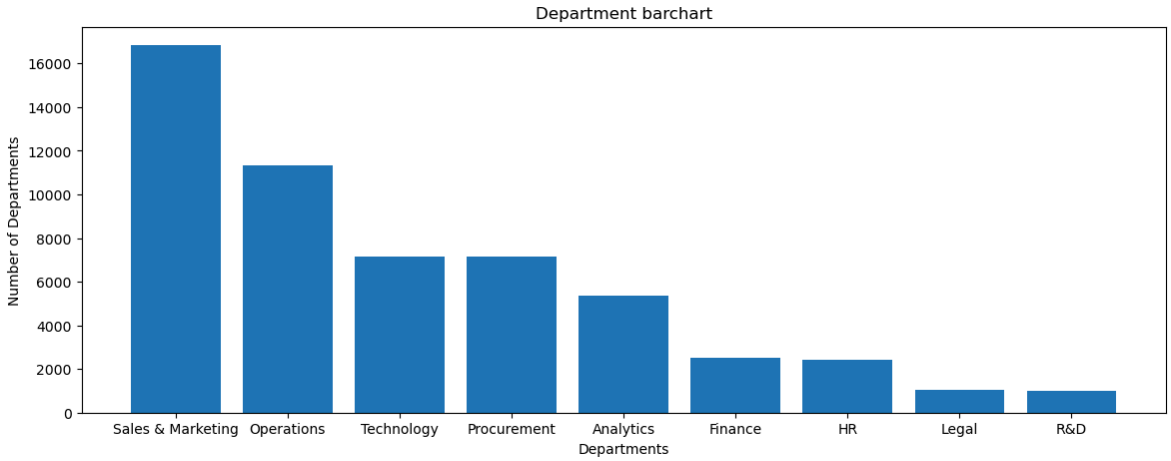
Out[8]:
```
pandas.core.series.Series
```

In [9]:
```
cdf=train_df['department'].value_counts()
keys=cdf.keys()
values=cdf.values
col=['Lables','Counts']
df=pd.DataFrame(zip(keys,values),columns=col)
df.to_csv('department.csv',index=False)
```

In [10]:
```
import matplotlib.pyplot as plt
plt.figure(figsize=(14,5))
```

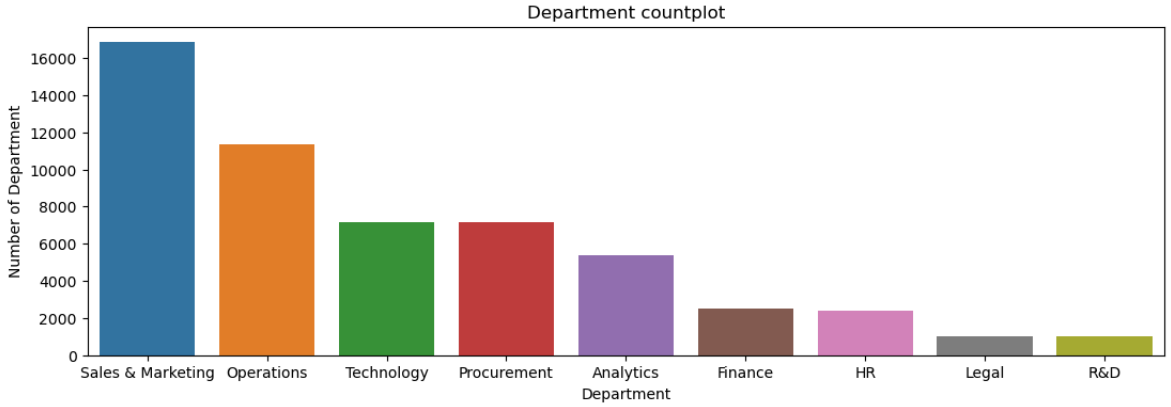
```
plt.bar('Lables','Counts',data=df)

plt.title('Department barchart')
plt.xlabel('Departments')
plt.ylabel('Number of Departments')
plt.show()
```




In [15]:
```
import seaborn as sns
plt.figure(figsize=(13,4))
order=['Sales & Marketing','Operations','Technology','Procurement','Analytics','
sns.countplot(data=train_df,x='department',order=order)

plt.title('Department countplot')
plt.xlabel('Department')
plt.ylabel('Number of Department')
plt.savefig('train_df.jpg')
plt.show()
```



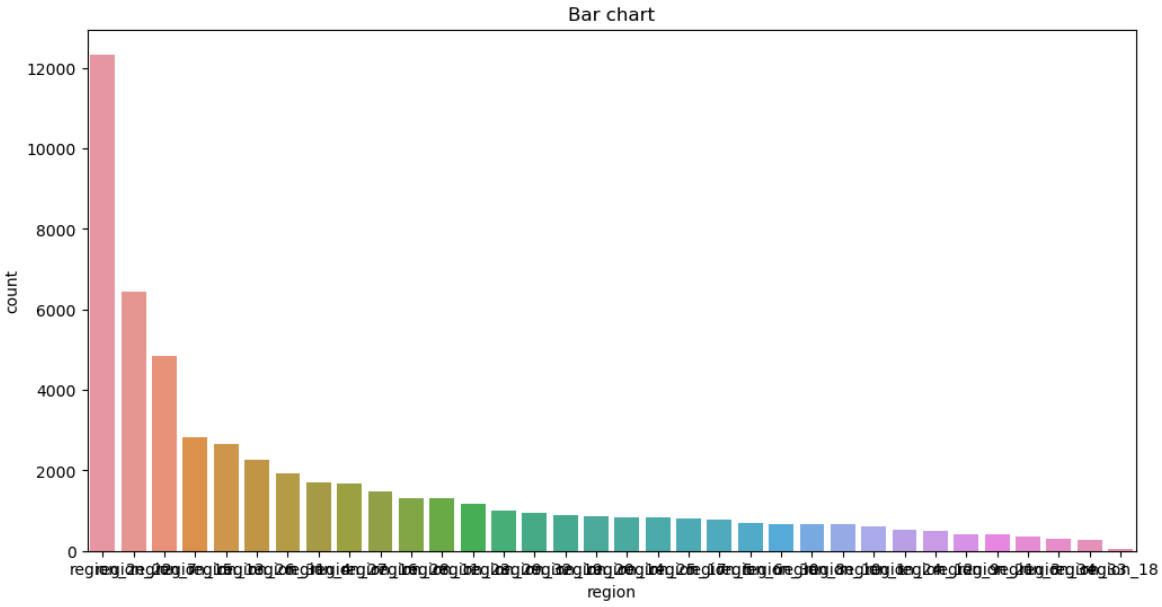

In [16]:
```
import os
os.getcwd()
```

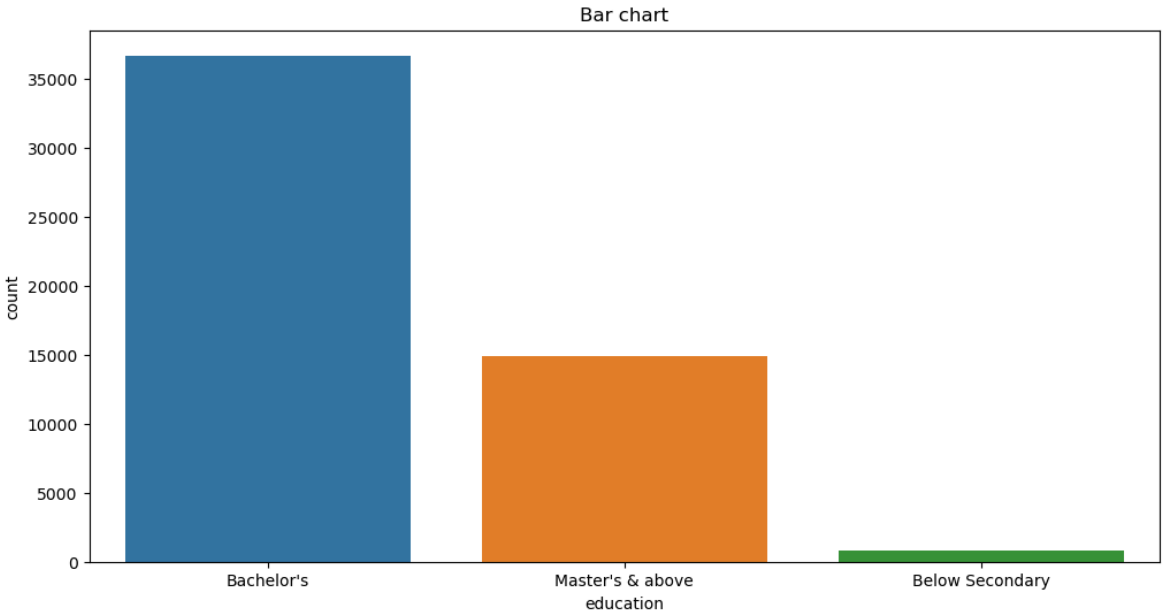
Out[16]:
```
'C:\\Users\\DELL\\Documents\\project'
```

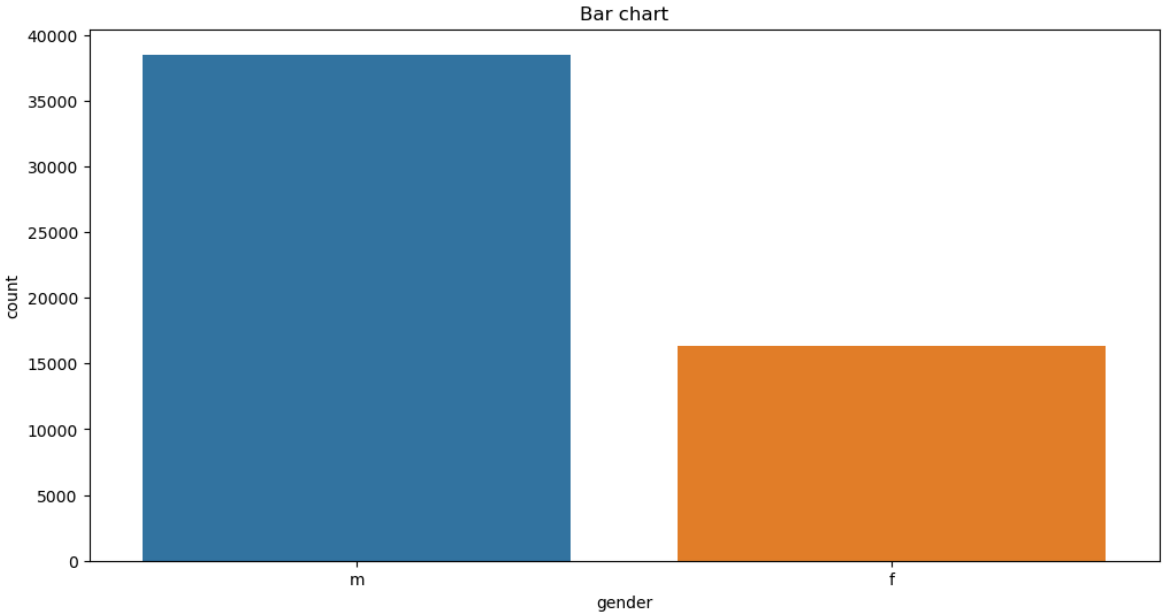
In [20]:
```
try:
    root_directory=os.getcwd()
    new_folder='Train_df'
    new_dir=os.path.join(root_directory,new_folder)
    os.makedirs(new_dir)
except Exception as e:
    print(e)
```

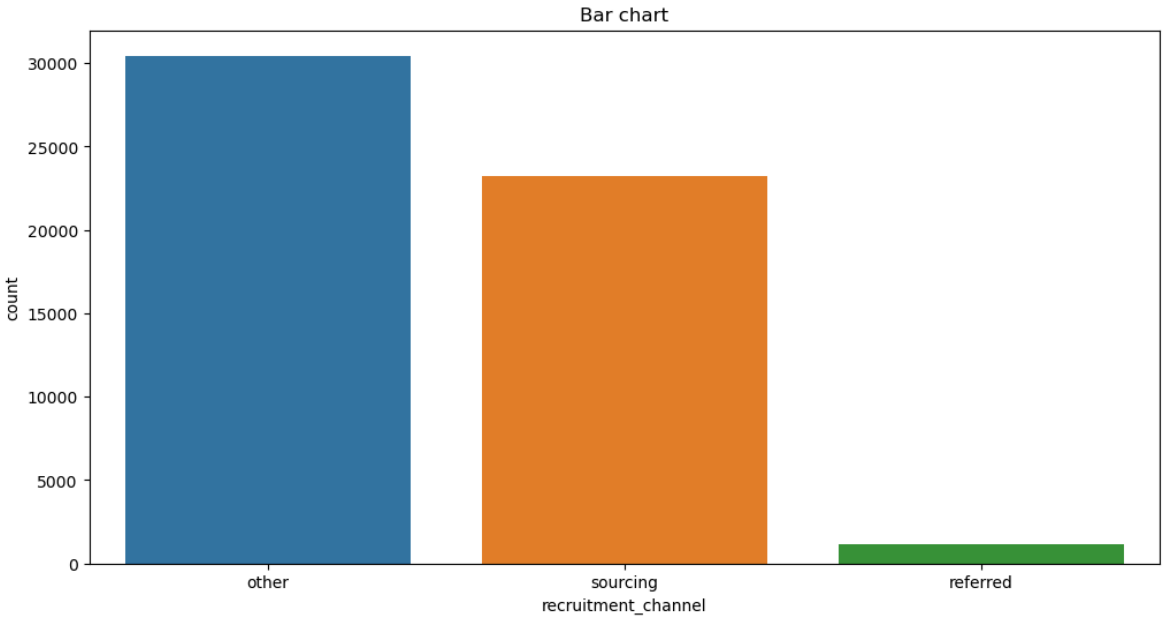
```
[WinError 183] Cannot create a file when that file already exists: 'C:\\Users\\DELL\\Documents\\project\\Train_df'
```

In [31]:
```
import seaborn as sns
for i in cat_columns[1:]:
    plt.figure(figsize=(12,6))
    order_continents=train_df[i].value_counts().keys()
    sns.countplot(data=train_df,x=i,order=order_continents)
    plt.title('Bar chart')
    plt.savefig(f'{new_dir}\\{i}_seaborn.jpg')
    plt.show()
```






In [42]:
```
for i in cat_columns[1:]:
    cdf=train_df[i].value_counts()
    keys=cdf.keys()
    values=cdf.values
    cols=['Lables','Counts']
    df=pd.DataFrame(zip(keys,values),columns=cols)
    df.to_csv(f'{i}.csv',index=False)
```

In [47]:
```
import os
os.getcwd()
try:
    root_directory=os.getcwd()
    new_folder='Train_df2'
    new_dir=os.path.join(root_directory,new_folder)
    os.makedirs(new_dir)
except Exception as e:
    print(e)
```

```
[WinError 183] Cannot create a file when that file already exists: 'C:\\Users\\DE
LL\\Documents\\project\\Train_df2'
```

In [49]:
```
dfs=os.listdir(r"C:\Users\DELL\Documents\project\Train_df2")
dfs
```

Out[49]:
```
['Education.csv', 'Gender.csv', 'recruitment_channel.csv', 'region.csv']
```

In [50]:
```
root_directory=os.getcwd()
new_folder='Train_df2'
dir=os.path.join(root_directory,new_folder)
dir
```

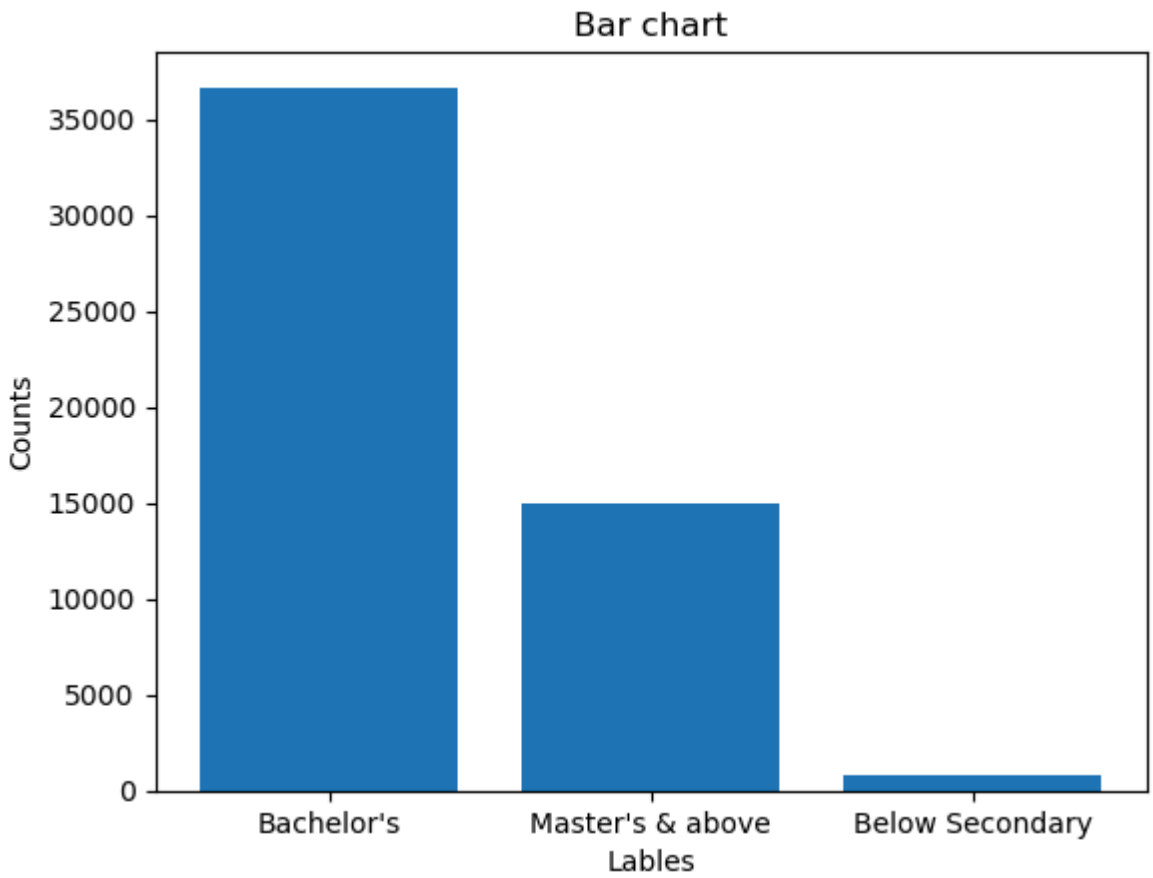
Out[50]:
```
'C:\\Users\\DELL\\Documents\\project\\Train_df2'
```

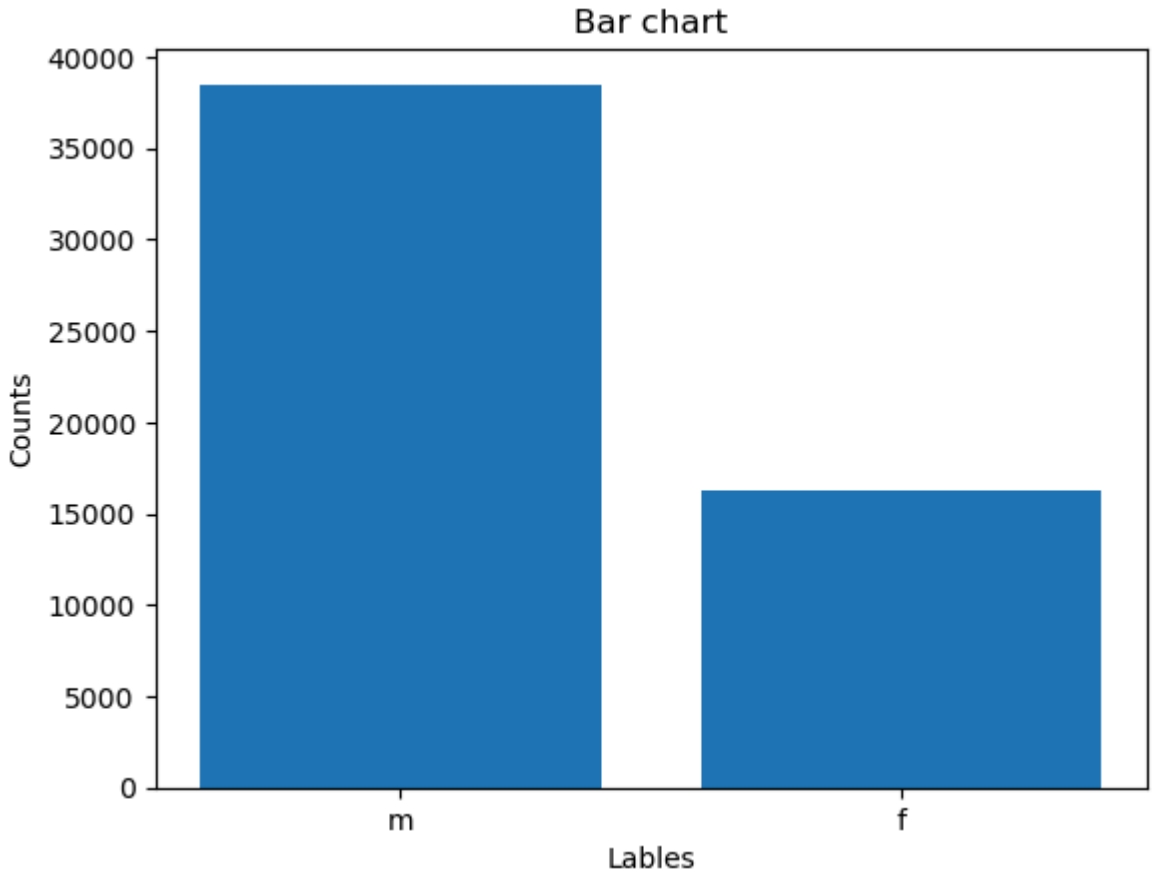
In [51]:
```
dfs=os.listdir(dir)
dfs
```

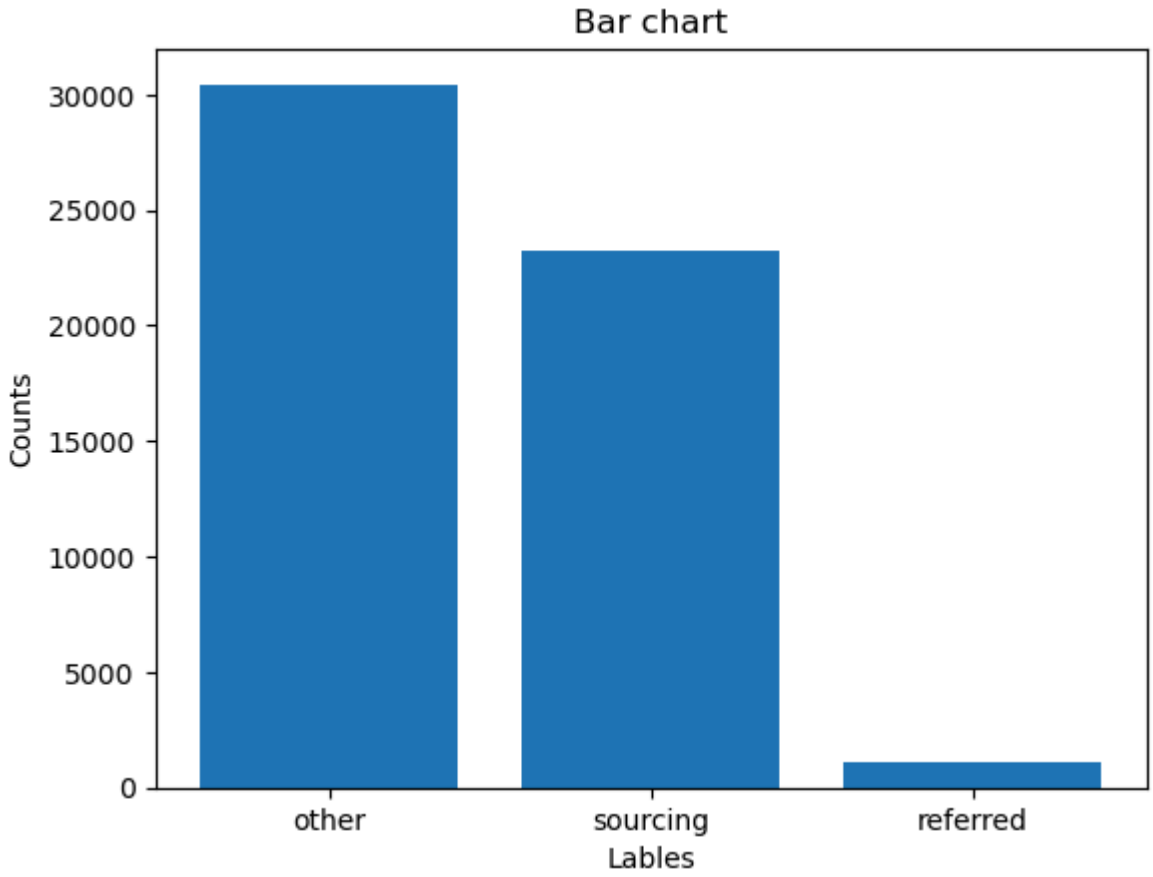
Out[51]:
```
['Education.csv', 'Gender.csv', 'recruitment_channel.csv', 'region.csv']
```

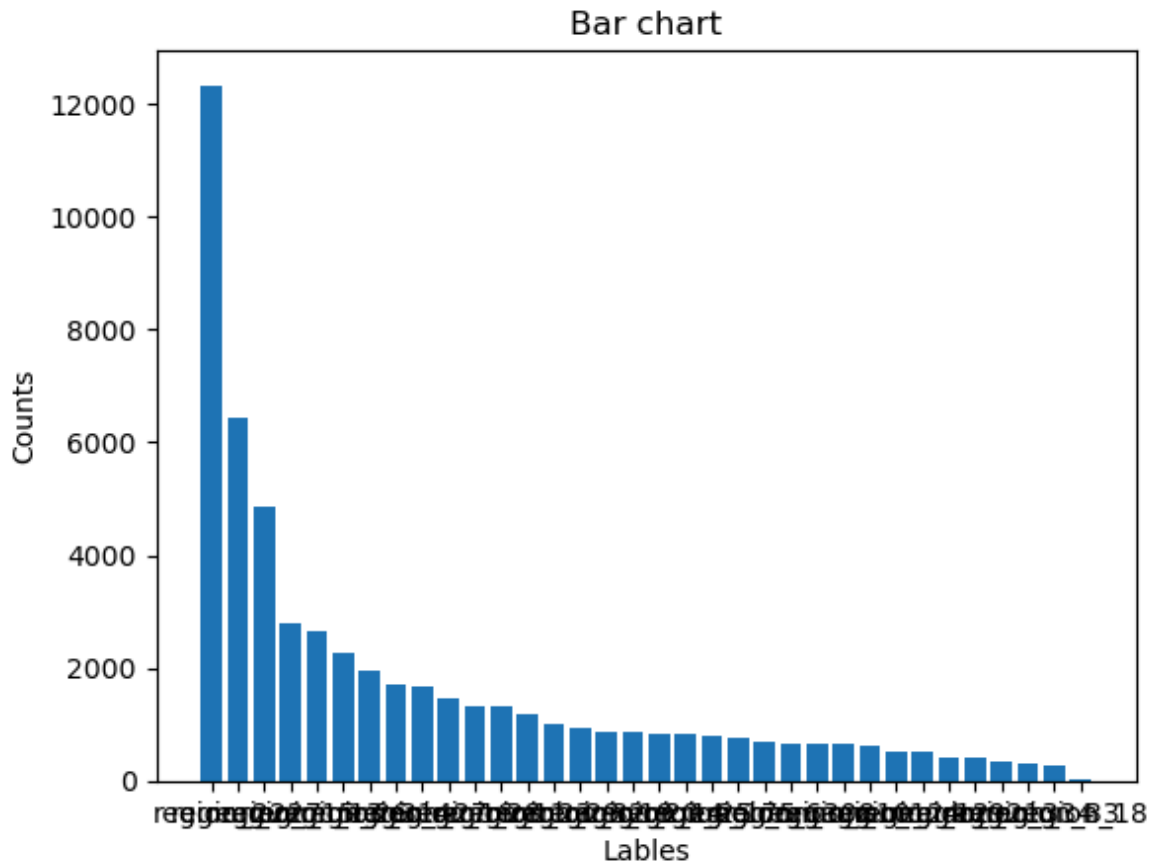
In [54]:
```
import matplotlib.pyplot as plt
for i in dfs:
    df=pd.read_csv(f'{dir}\\{i}')
    plt.bar('Lables','Counts',data=df)

    plt.title('Bar chart')
    plt.xlabel('Lables')
    plt.ylabel('Counts')
    plt.savefig(f'{new_dir}\\{i}_matplotlip.jpg')
    plt.show()
```

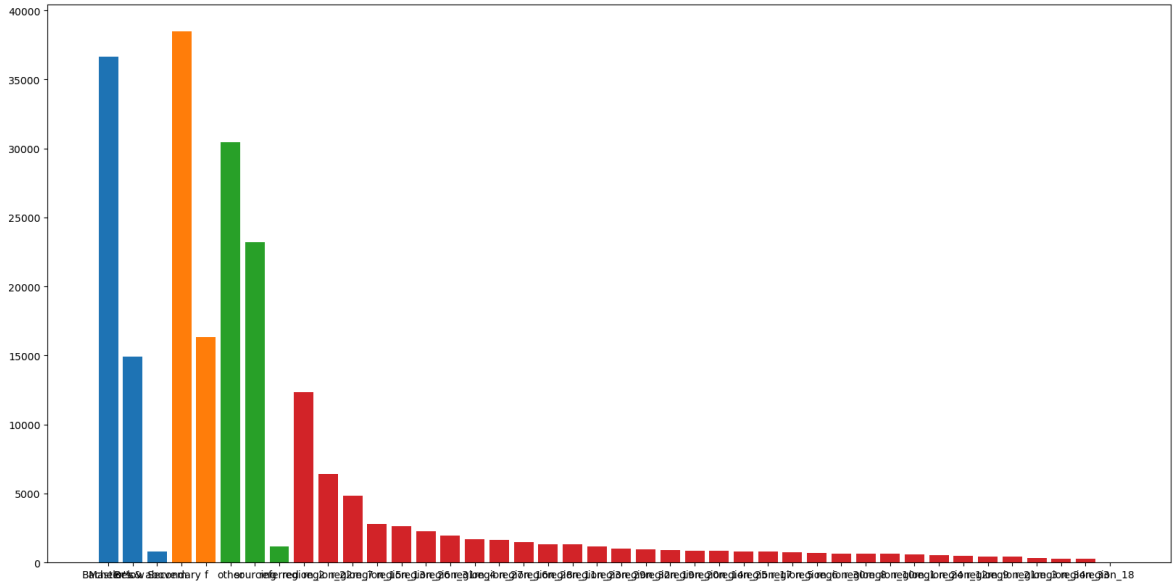



Bar chart
40000
35000
30000
25000
20000
15000
10000
5000
0
Counts
m
f
Lables

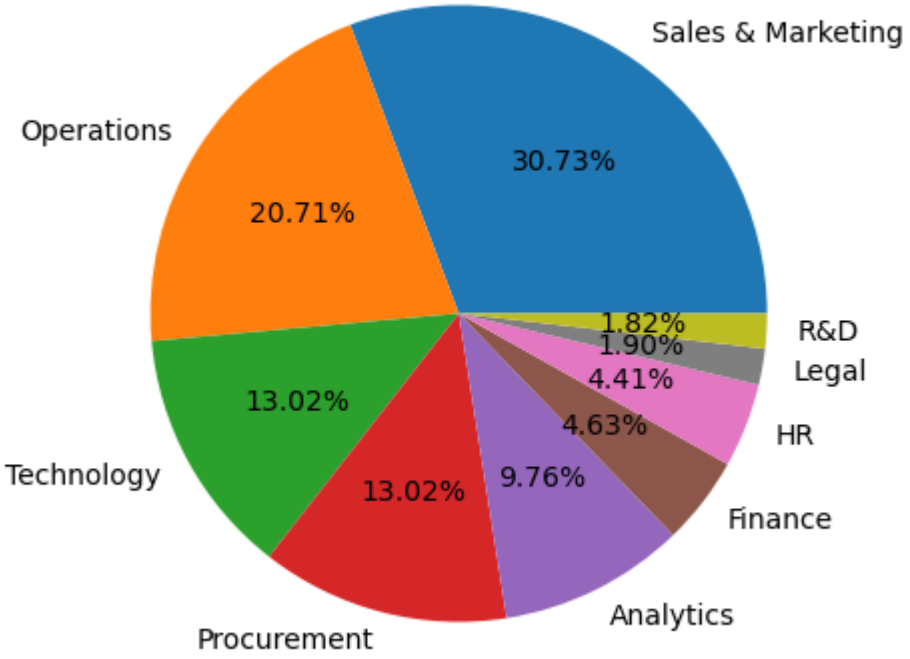

Bar chart
30000
25000
20000
15000
10000
5000
0
Counts
other
sourcing
referred
Lables



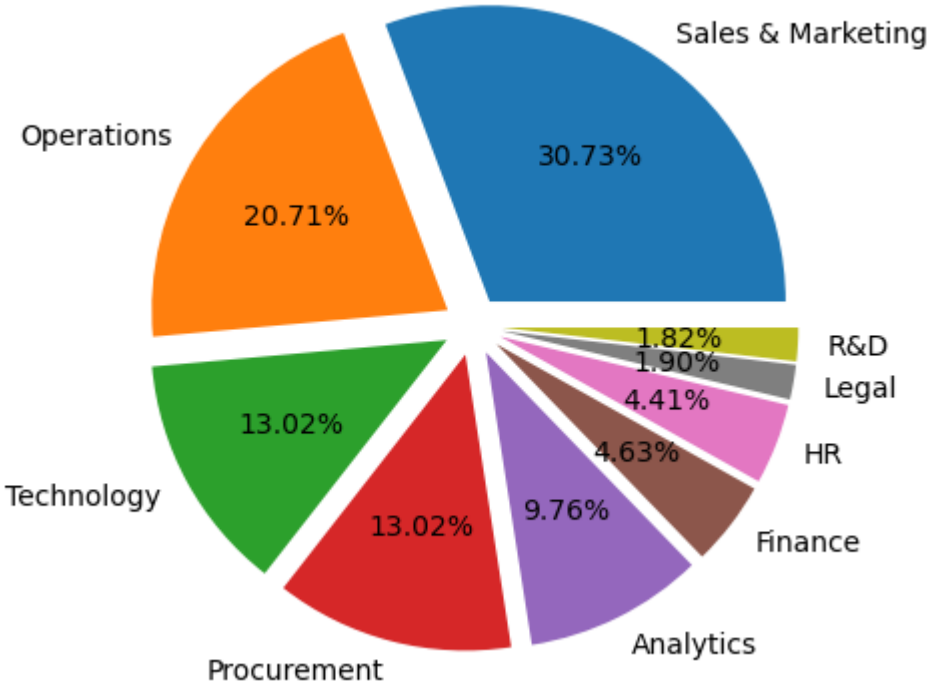

In [63]:
```
import matplotlib.pyplot as plt
plt.figure(figsize=(20,10))
for i in dfs:
    df=pd.read_csv(f'{dir}\\{i}')
    plt.bar('Lables','Counts',data=df)
```




In [71]:
```
import matplotlib.pyplot as plt
plt.figure(figsize=(10,5))
keys=train_df['department'].value_counts().keys()
values=train_df['department'].value_counts().values
plt.pie(values,labels=keys,autopct='%0.2f%%')
plt.show()
```




In [75]:
```
import matplotlib.pyplot as plt
keys=train_df['department'].value_counts().keys()
values=train_df['department'].value_counts().values
plt.pie(values,labels=keys,explode=[0.1]*len(keys),autopct='%0.2f%%')
plt.show()
```




In [ ]:
```
import matplotlib.pyplot as plt
keys=train_df['department'].value_counts().keys()
values=train_df['department'].value_counts().values
plt.pie(values,lables=keys,explode=[0.1]*len(keys),autopct='%0.2f%%',startangle=
```