0. Pandas Tutorials

April 3, 2023

[ ]:
```
import pandas as pd
```

[ ]:
```
df = pd.read_excel(io="i3student.xlsx", index_col=0)
df.head()
```

[ ]:

No	Student ID	Student Name	Gender	Department	Email
1	e20200861	BUT CHEABLENG	M	I3AMS	e20200861@itc.edu.kh
2	e20201690	BUTH KHEMRA	M	I3AMS	e20201690@itc.edu.kh
3	e20201131	CHEA MAKARA	M	I3AMS	e20201131@itc.edu.kh
4	e20200702	CHEA ROTHA	M	I3AMS	e20200702@itc.edu.kh
5	e20200934	CHHON CHAINA	M	I3AMS	e20200934@itc.edu.kh

[ ]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1385 entries, 1 to 1385
Data columns (total 5 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   Student ID    1385 non-null   object
 1   Student Name  1385 non-null   object
 2   Gender        1385 non-null   object
 3   Department    1385 non-null   object
 4   Email         1385 non-null   object
dtypes: object(5)
memory usage: 64.9+ KB
```

[ ]:
```
df.describe()
```

[ ]:

	Student ID	Student Name	Gender	Department	Email
count	1385	1385	1385	1385	1385
unique	1385	1384	2	11	1385
top	e20200861	SOK RATHA	M	I3GCI	e20200861@itc.edu.kh
freq	1	2	947	219	1

[ ]:
```
sok_ratha = df.loc[df.loc[:, 'Student Name'] == 'SOK RATHA', :]
print(sok_ratha)
```

```
     Student ID Student Name Gender Department                 Email
No
758   e20201026    SOK RATHA      M      I3GEE  e20201026@itc.edu.kh
1091  e20200567    SOK RATHA      M      I3GIM  e20200567@itc.edu.kh
```

[ ]:
```
ams_selection = df.loc[:, 'Department'] == 'I3AMS'
i3ams = df.loc[ams_selection, :]
print(i3ams)
```

```
    Student ID     Student Name Gender Department                 Email
No
1    e20200861    BUT CHEABLENG      M      I3AMS  e20200861@itc.edu.kh
2    e20201690      BUTH KHEMRA      M      I3AMS  e20201690@itc.edu.kh
3    e20201131      CHEA MAKARA      M      I3AMS  e20201131@itc.edu.kh
4    e20200702       CHEA ROTHA      M      I3AMS  e20200702@itc.edu.kh
5    e20200934     CHHON CHAINA      M      I3AMS  e20200934@itc.edu.kh
..         …              …    …        …                   …
84   e20201068    VINLAY ANUSAR      M      I3AMS  e20201068@itc.edu.kh
85   e20200745         YA MANON      M      I3AMS  e20200745@itc.edu.kh
86   e20200625  YIT BUN VATHANA      M      I3AMS  e20200625@itc.edu.kh
87   e20200727     YOU PHAKKORN      M      I3AMS  e20200727@itc.edu.kh
301  e20200370        PHY KANHA      F      I3AMS  e20200370@itc.edu.kh

[88 rows x 5 columns]
```

[ ]:
```
df_sorted = df.sort_values(by=['Student Name', 'Gender'])
print(df_sorted)
```

```
     Student ID     Student Name Gender Department                 Email
No
635   e20201044      AING NARONG      M      I3GEE  e20201044@itc.edu.kh
416   e20200937     AM CHANPANHA      M      I3GCI  e20200937@itc.edu.kh
417   e20190006       AN BRONITH      M      I3GCI  e20190006@itc.edu.kh
184   e20200429        AN CHANNA      F      I3GCA  e20200429@itc.edu.kh
968   e20200357          AN RITHY      M      I3GIM  e20200357@itc.edu.kh
…         …              …    …        …                   …
1129  e20200595    YOUVARA DANIN      M      I3GIM  e20200595@itc.edu.kh
1242  e20201462  YUN DARAVATHANA      M      I3GRU  e20201462@itc.edu.kh
634   e20201509       YUN VANNET      M      I3GCI  e20201509@itc.edu.kh
805   e20200035       YUN YISEAN      M      I3GEE  e20200035@itc.edu.kh
415   e20201013       YUOS SOKNY      M      I3GCA  e20201013@itc.edu.kh

[1385 rows x 5 columns]
```

[ ]:
```
df_sliced = df.loc[:, ['Department', 'Gender']]
print(df_sliced)
```

```
     Department Gender
No
```

```
1          I3AMS      M
2          I3AMS      M
3          I3AMS      M
4          I3AMS      M
5          I3AMS      M
…          …    …
1381       I3GTR      M
1382       I3GTR      M
1383       I3GTR      F
1384       I3GTR      M
1385       I3GTR      M

[1385 rows x 2 columns]
```

[ ]:
```
df_grouped = df_sliced.groupby(by=['Department', 'Gender']).value_counts().
  ↪reset_index()
print(df_grouped)
```

```
   Department Gender    0
0       I3AMS      F   22
1       I3AMS      M   66
2       I3GAR      F   37
3       I3GAR      M   60
4       I3GCA      F  134
5       I3GCA      M   58
6       I3GCI      F   19
7       I3GCI      M  200
8       I3GEE      F   38
9       I3GEE      M  140
10      I3GGG      F   35
11      I3GGG      M   59
12      I3GIC      F   20
13      I3GIC      M   60
14      I3GIM      F   26
15      I3GIM      M  139
16      I3GRU      F   56
17      I3GRU      M   66
18      I3GTI      F   28
19      I3GTI      M   55
20      I3GTR      F   23
21      I3GTR      M   44
```

[ ]:
```
df_renamed = df_grouped.rename(columns={0: 'Count'})
print(df_renamed)
```

```
   Department Gender  Count
0       I3AMS      F     22
1       I3AMS      M     66
2       I3GAR      F     37
```

```
3       I3GAR      M     60
4       I3GCA      F    134
5       I3GCA      M     58
6       I3GCI      F     19
7       I3GCI      M    200
8       I3GEE      F     38
9       I3GEE      M    140
10      I3GGG      F     35
11      I3GGG      M     59
12      I3GIC      F     20
13      I3GIC      M     60
14      I3GIM      F     26
15      I3GIM      M    139
16      I3GRU      F     56
17      I3GRU      M     66
18      I3GTI      F     28
19      I3GTI      M     55
20      I3GTR      F     23
21      I3GTR      M     44
```

[ ]:
```
df_pivoted = df_renamed.pivot_table(index='Department', columns='Gender',␣
 ↪values='Count')
print(df_pivoted)
```

```
Gender        F    M
Department
I3AMS        22   66
I3GAR        37   60
I3GCA       134   58
I3GCI        19  200
I3GEE        38  140
I3GGG        35   59
I3GIC        20   60
I3GIM        26  139
I3GRU        56   66
I3GTI        28   55
I3GTR        23   44
```

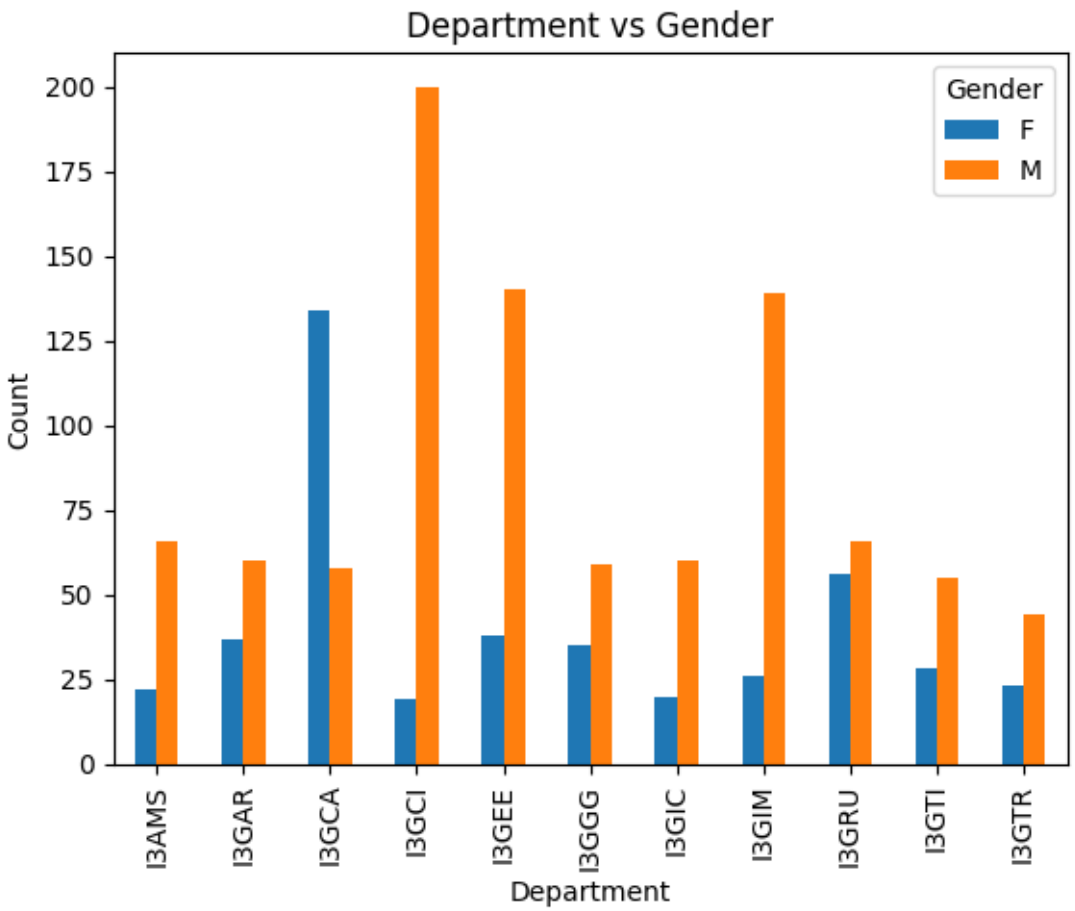
[ ]:
```
# print(pd.melt.__doc__)
df_reindex = df_pivoted.reset_index()
df_melt = df_reindex.melt(id_vars=['Department'], value_vars=['F', 'M'],␣
 ↪value_name='Count')
df_melt
```

[ ]:
```
   Department Gender  Count
0       I3AMS      F     22
1       I3GAR      F     37
2       I3GCA      F    134
```

```
3       I3GCI      F     19
4       I3GEE      F     38
5       I3GGG      F     35
6       I3GIC      F     20
7       I3GIM      F     26
8       I3GRU      F     56
9       I3GTI      F     28
10      I3GTR      F     23
11      I3AMS      M     66
12      I3GAR      M     60
13      I3GCA      M     58
14      I3GCI      M    200
15      I3GEE      M    140
16      I3GGG      M     59
17      I3GIC      M     60
18      I3GIM      M    139
19      I3GRU      M     66
20      I3GTI      M     55
21      I3GTR      M     44
```

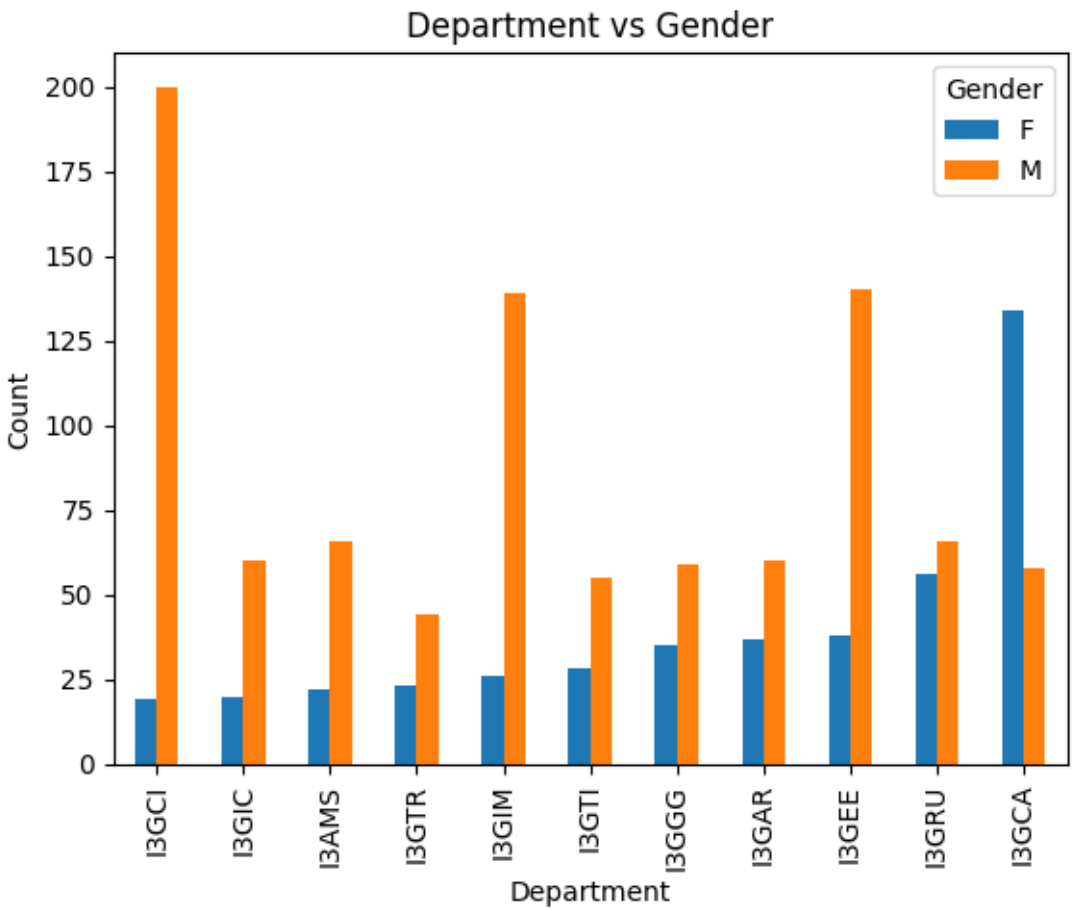
[ ]:
```
df_pivoted.plot(kind='bar', title='Department vs Gender', ylabel='Count')
```

[ ]:
```
<AxesSubplot: title={'center': 'Department vs Gender'}, xlabel='Department',
ylabel='Count'>
```



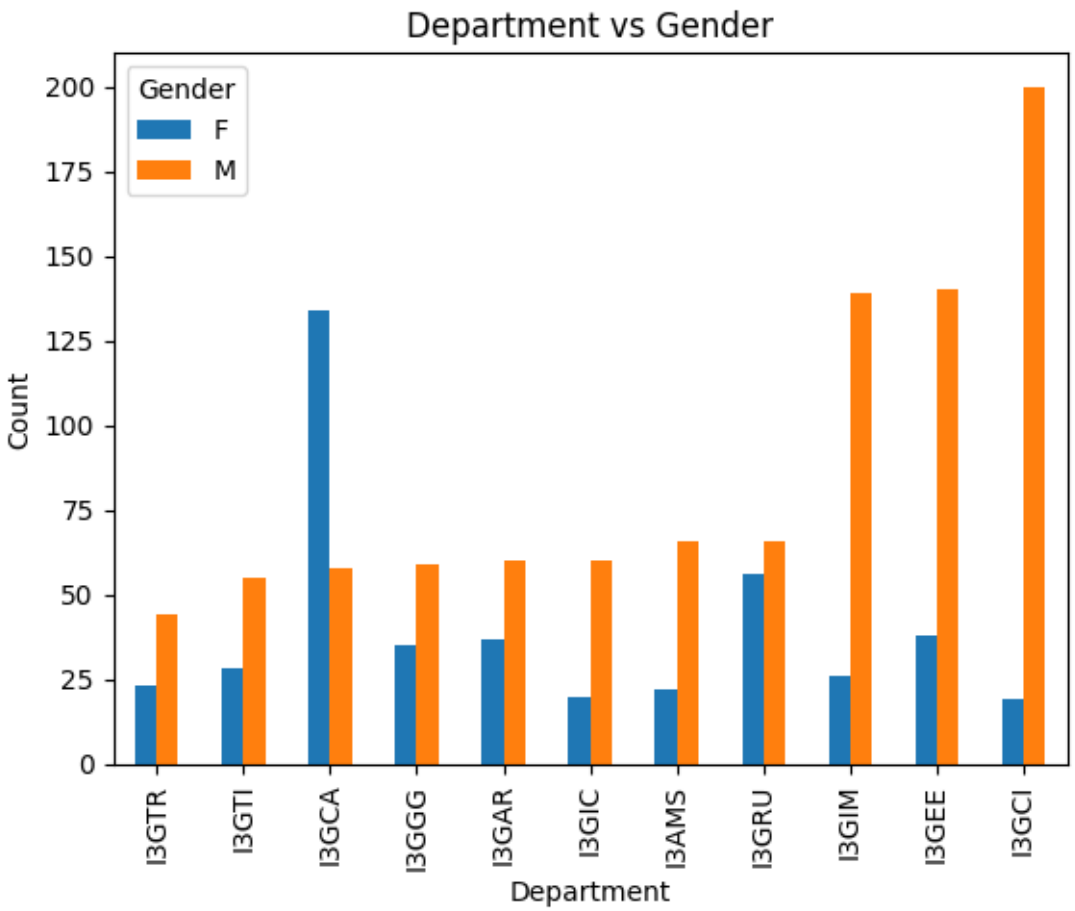
[ ]:
```
df_pivoted.sort_values(by=['F']).plot(kind='bar', title='Department vs Gender',␣
↪ylabel='Count')
```

[ ]:
```
<AxesSubplot: title={'center': 'Department vs Gender'}, xlabel='Department',
ylabel='Count'>
```



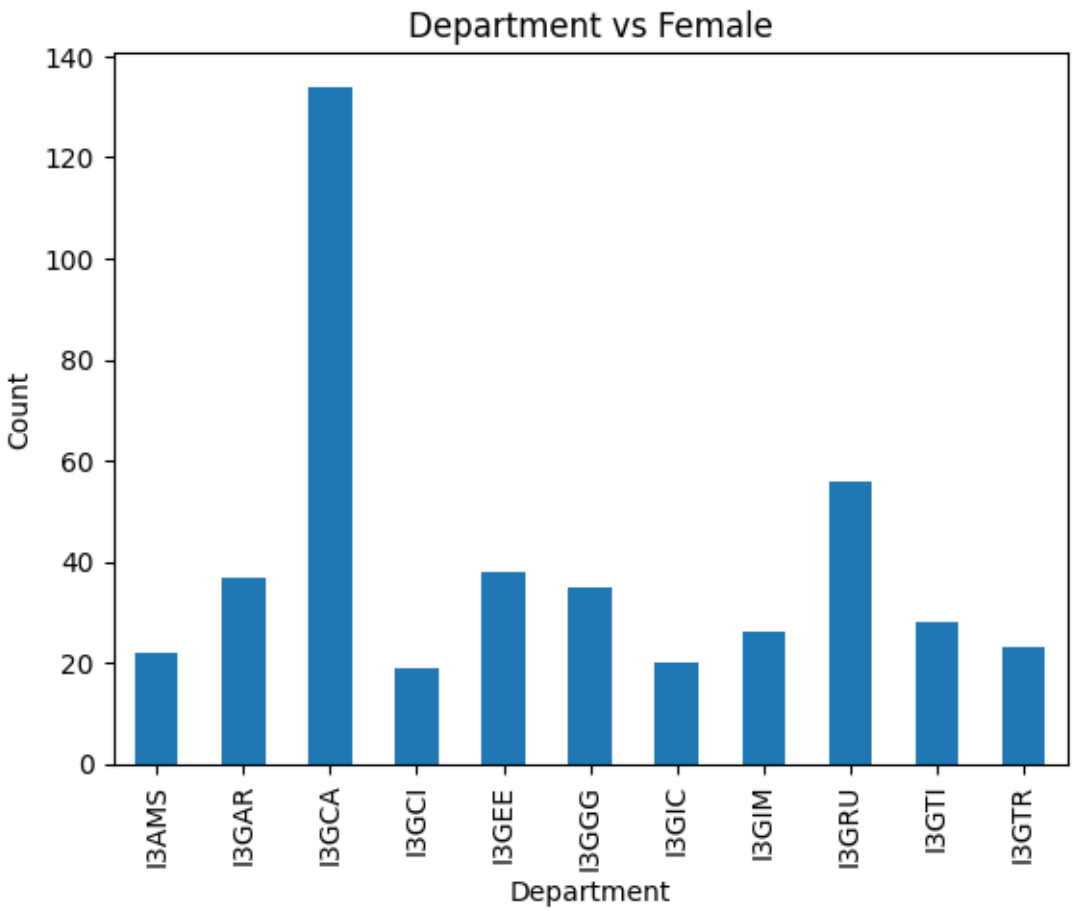
[ ]:
```
df_pivoted.sort_values(by=['M']).plot(kind='bar', title='Department vs Gender',␣
↪ylabel='Count')
```

[ ]:
```
<AxesSubplot: title={'center': 'Department vs Gender'}, xlabel='Department',
ylabel='Count'>
```



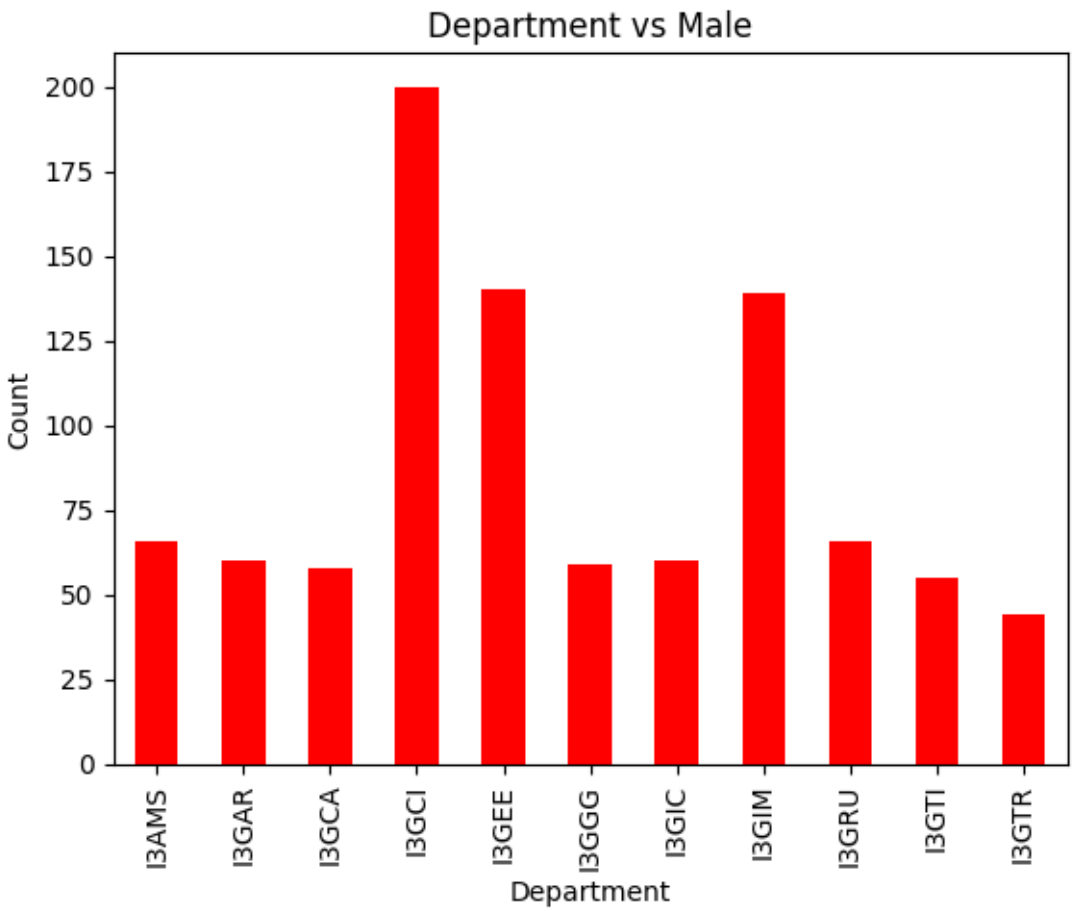
[ ]:
```
df_pivoted.loc[:, 'F'].plot(kind='bar', title='Department vs Female',␣
↪ylabel='Count')
```

[ ]:
```
<AxesSubplot: title={'center': 'Department vs Female'}, xlabel='Department',
ylabel='Count'>
```



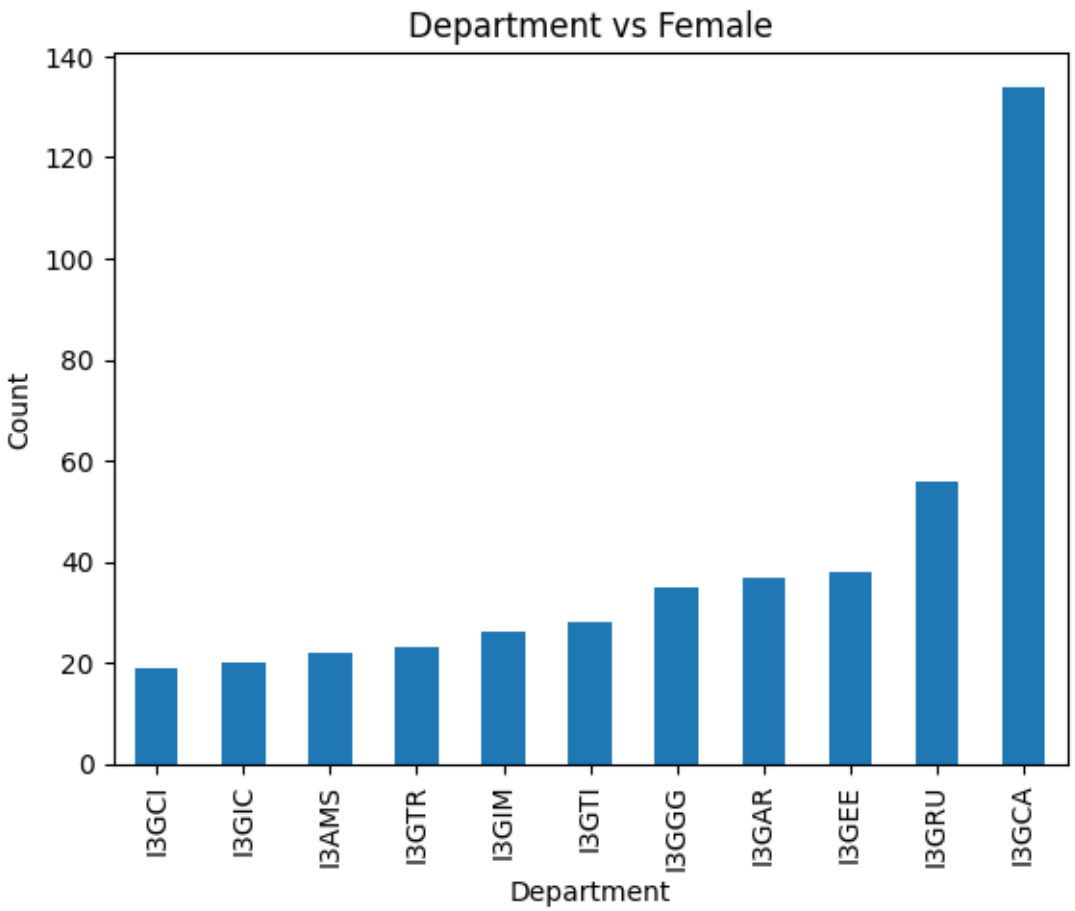
[ ]:
```
df_pivoted.loc[:, 'M'].plot(kind='bar', title='Department vs Male',␣
↪ylabel='Count', color='red')
```

[ ]:
```
<AxesSubplot: title={'center': 'Department vs Male'}, xlabel='Department',
ylabel='Count'>
```



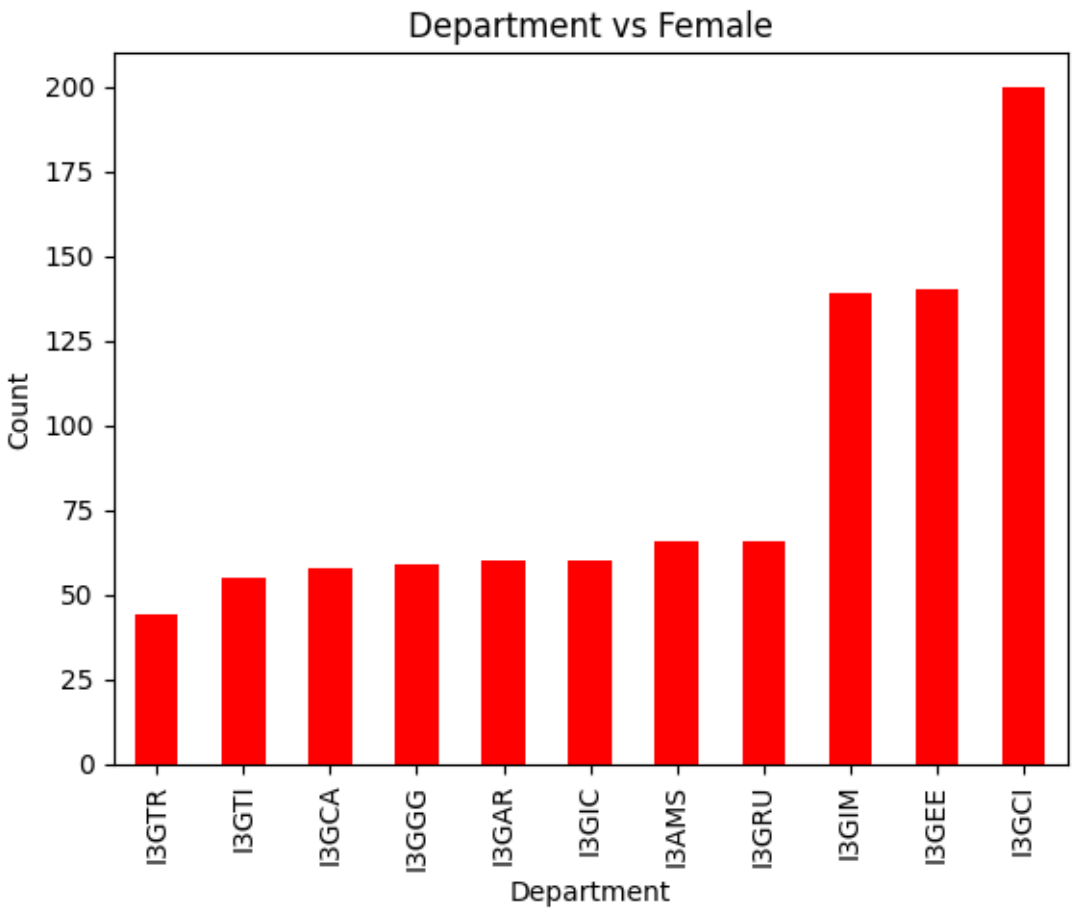
[ ]:
```
df_pivoted.loc[:, 'F'].sort_values(ascending=True).plot(kind='bar',␣
↪title='Department vs Female', ylabel='Count')
```

[ ]:
```
<AxesSubplot: title={'center': 'Department vs Female'}, xlabel='Department',
ylabel='Count'>
```



[ ]:
```
df_pivoted.loc[:, 'M'].sort_values(ascending=True).plot(kind='bar',␣
↪title='Department vs Female', ylabel='Count', color='red')
```

[ ]:
```
<AxesSubplot: title={'center': 'Department vs Female'}, xlabel='Department',
ylabel='Count'>
```



[ ]:
```
surname = df.loc[:, 'Student Name'].apply(func = lambda x: x.split(sep=" ")[0])
name = df.loc[:, 'Student Name'].apply(func = lambda x: x.split(sep=" ")[1])
print(df.columns)
df_new = df.copy()
df_new.insert(loc=2, column='Surname', value=surname)
df_new.insert(loc=3, column='Name', value=name)
print(df_new.head())
```

```
Index(['Student ID', 'Student Name', 'Gender', 'Department', 'Email'],
dtype='object')
   Student ID   Student Name Surname       Name Gender Department  \
No
1   e20200861  BUT CHEABLENG     BUT  CHEABLENG      M      I3AMS
2   e20201690    BUTH KHEMRA    BUTH     KHEMRA      M      I3AMS
3   e20201131    CHEA MAKARA    CHEA     MAKARA      M      I3AMS
4   e20200702     CHEA ROTHA    CHEA      ROTHA      M      I3AMS
5   e20200934   CHHON CHAINA   CHHON     CHAINA      M      I3AMS

                    Email
```

```
No
1   e20200861@itc.edu.kh
2   e20201690@itc.edu.kh
3   e20201131@itc.edu.kh
4   e20200702@itc.edu.kh
5   e20200934@itc.edu.kh
```

[ ]:
```
df_new = df_new.drop(columns='Student Name')
print(df_new.head())
```

```
   Student ID Surname       Name Gender Department                 Email
No
1   e20200861     BUT  CHEABLENG      M      I3AMS  e20200861@itc.edu.kh
2   e20201690    BUTH     KHEMRA      M      I3AMS  e20201690@itc.edu.kh
3   e20201131    CHEA     MAKARA      M      I3AMS  e20201131@itc.edu.kh
4   e20200702    CHEA      ROTHA      M      I3AMS  e20200702@itc.edu.kh
5   e20200934   CHHON     CHAINA      M      I3AMS  e20200934@itc.edu.kh
```

[ ]:
```
df_new.loc[:, 'Department'] = df_new.loc[:, 'Department'].apply(func = lambda x:
  ↪ x[2:])
print(df_new.head())
```

```
   Student ID Surname       Name Gender Department                 Email
No
1   e20200861     BUT  CHEABLENG      M        AMS  e20200861@itc.edu.kh
2   e20201690    BUTH     KHEMRA      M        AMS  e20201690@itc.edu.kh
3   e20201131    CHEA     MAKARA      M        AMS  e20201131@itc.edu.kh
4   e20200702    CHEA      ROTHA      M        AMS  e20200702@itc.edu.kh
5   e20200934   CHHON     CHAINA      M        AMS  e20200934@itc.edu.kh
```

[ ]:
```
def Gender(g: str) -> str:
    if g.upper() == 'M':
        gender = 'Male'
    elif g.upper() == "F":
        gender = 'Female'
    else:
        gender = g
    return gender


df_new.loc[:, 'Gender'] = df_new.loc[:, 'Gender'].apply(func=Gender)
print(df_new.head())
```

```
   Student ID Surname       Name Gender Department                 Email
No
1   e20200861     BUT  CHEABLENG   Male        AMS  e20200861@itc.edu.kh
2   e20201690    BUTH     KHEMRA   Male        AMS  e20201690@itc.edu.kh
3   e20201131    CHEA     MAKARA   Male        AMS  e20201131@itc.edu.kh
4   e20200702    CHEA      ROTHA   Male        AMS  e20200702@itc.edu.kh
```

```
5   e20200934   CHHON     CHAINA   Male        AMS  e20200934@itc.edu.kh
```

[ ]:
```
select_007 = df_new.loc[:,'Student ID'].str.endswith("007")
id_007 = df_new.loc[select_007,:]
print(id_007)
```

```
     Student ID Surname     Name  Gender Department                 Email
No
307   e20220007    PROM    SEREY  Female        GCA  e20220007@itc.edu.kh
1318  e20200007     THY  SOKLEAB    Male        GTI  e20200007@itc.edu.kh
```

[ ]:
```
select_2022 = df_new.loc[:, 'Student ID'].str.startswith("e2022")
id_2022 = df_new.loc[select_2022, :]
print(id_2022)
```

```
     Student ID  Surname           Name  Gender Department  \
No
189   e20220001     CHEA          SOKLY    Male        GCA
192   e20220002   CHHORN           SINA  Female        GCA
229   e20220003   KHEANG          SOKLY  Female        GCA
254   e20220004    LEANG       SOKVISAL    Male        GCA
256   e20220005      LIM          PISEY  Female        GCA
304   e20220006      POV         PHARUM  Female        GCA
307   e20220007     PROM          SEREY  Female        GCA
329   e20220008     SEAB          THIDA  Female        GCA
385   e20220009   TOEUNG          CHAIM    Male        GCA
520   e20220010      NAP          DANET  Female        GCI
602   e20220011  SROEUNG        KALLYAN  Female        GCI
714   e20220012       NY       SREYLEAK  Female        GEE
777   e20220013     SROS  SOKSOPHEAKTRA    Male        GEE
1000  e20220014     HOUN      MENGHEANG    Male        GIM
1068  e20220015      RIN       VENGTHAI    Male        GIM

                     Email
No
189   e20220001@itc.edu.kh
192   e20220002@itc.edu.kh
229   e20220003@itc.edu.kh
254   e20220004@itc.edu.kh
256   e20220005@itc.edu.kh
304   e20220006@itc.edu.kh
307   e20220007@itc.edu.kh
329   e20220008@itc.edu.kh
385   e20220009@itc.edu.kh
520   e20220010@itc.edu.kh
602   e20220011@itc.edu.kh
714   e20220012@itc.edu.kh
777   e20220013@itc.edu.kh
1000  e20220014@itc.edu.kh
```

```
1068  e20220015@itc.edu.kh
```

[ ]:
```
id_e2022 = df_new.loc[:, 'Student ID'].str.startswith("e2022")
dep_gca = df_new.loc[:, 'Department'] == 'GCA'
gen_fem = df_new.loc[:, 'Gender'] == 'Female'
query = id_e2022 & dep_gca & gen_fem
result = df_new.loc[query, 'Student ID':'Name']
print(result)
```

```
    Student ID Surname    Name
No
192  e20220002  CHHORN    SINA
229  e20220003  KHEANG   SOKLY
256  e20220005     LIM   PISEY
304  e20220006     POV  PHARUM
307  e20220007    PROM   SEREY
329  e20220008    SEAB   THIDA
```

[ ]:
```
dep_ams = df_new.loc[:, "Department"] == "AMS"
id_e2020 = df_new.loc[:, "Student ID"].str.startswith("e2020")
surname_start_c = df_new.loc[:, "Surname"].str.startswith("C")
name_end_a = df_new.loc[:, "Name"].str.endswith("A")
query = dep_ams & id_e2020 & surname_start_c & name_end_a
result = df_new.loc[query, ['Surname', 'Name', 'Department']].
  ↪sort_values(by='Name')
print(result)
```

```
   Surname        Name Department
No
5    CHHON      CHAINA        AMS
6    CHORN  CHANLAKHNA        AMS
3     CHEA      MAKARA        AMS
4     CHEA       ROTHA        AMS
```

[ ]:
```
with pd.ExcelWriter(path='i3student_clean.xlsx', mode='w') as writer:
    df_new.to_excel(excel_writer=writer)
```