ศุภิสรา อุดมลาภ รหัสนักศึกษา 6420412013

Data Journey

Dataset : มีทั้งหมด 2 ไฟล์ คือ ไฟล์จำนวนผู้มีงานทำจำแนกตามระดับการศึกษาที่สำเร็จ และ ไฟล์จำนวนผู้มีงานทำจำแนกตามอายุและเพศ

- ไฟล์จำนวนผู้มีงานทำจำแนกตามระดับการศึกษาที่สำเร็จ มีทั้งหมด 6499 แถวและ 8 คอลัมน์

ซึ่งคอลัมน์มี ปี,ไตรมาส,ภาค,พื้นที่,ระดับการศึกษา,จำนวนคน,หน่วย(พันคน),แหล่งที่มา

- ไฟล์จำนวนผู้มีงานทำจำแนกตามอายุและเพศ มีทั้งหมด 5832 แถวและ 8 คอลัมน์

ซึ่งคอลัมน์มี ปี,ไตรมาส,ภาค,ช่วงอายุ,เพศ,จำนวนคน,หน่วย(พันคน),แหล่งที่มา

แหล่งที่มาของข้อมูล คือ สำนักงานสถิติแห่งชาติ หรือที่เว็บไซต์ https://data.go.th/th/dataset/os_02_00002

Cleaning data : จากการตรวจสอบโดยใช้คำสั่ง df.isnull().sum() ทั้งสองไฟล์ พบว่า ไม่มีค่า Nanทุกตัวแปร

```
[ ]  1 df_worker.isnull().sum()
```

```
year           0
quarter        0
region         0
area           0
level_of_edu   0
value          0
unit           0
source         0
dtype: int64
```

```
  1 df_worker['region'].isna()
```

```
0       False
1       False
2       False
3       False
4       False
        ...
6494    False
6495    False
6496    False
6497    False
6498    False
Name: region, Length: 6499, dtype: bool
```

```
[ ]  1 df_worker['level_of_edu'].isna()
```

```
0       False
1       False
2       False
3       False
4       False
        ...
6494    False
6495    False
6496    False
6497    False
6498    False
Name: level_of_edu, Length: 6499, dtype: bool
```

```
[ ]  1 df_worker['region'].unique()
```

```
array(['ทั่วประเทศ', 'กรุงเทพมหานคร', 'ภาคกลาง', 'ภาคเหนือ',
       'ภาคตะวันออกเฉียงเหนือ', 'ภาคใต้'], dtype=object)
```

```
  1 df_worker['region'].value_counts()
```

```
ทั่วประเทศ                1247
ภาคกลาง                   1243
ภาคเหนือ                  1238
ภาคใต้                    1228
ภาคตะวันออกเฉียงเหนือ     1167
กรุงเทพมหานคร              376
Name: region, dtype: int64
```

```
[ ]  1 df_worker['level_of_edu'].unique()
```

```
array(['รวม', 'ไม่มีการศึกษา', 'ต่ำกว่าประถมศึกษา', 'ประถมศึกษา',
       'มัธยมศึกษาตอนต้น', 'มัธยมศึกษาตอนปลาย (สายสามัญ)',
       'มัธยมศึกษาตอนปลาย (สายอาชีวศึกษา)',
       'มัธยมศึกษาตอนปลาย (สายวิชาการศึกษา)', 'อุดมศึกษา (สายวิชาการ)',
       'อุดมศึกษา (สายวิชาชีพ)', 'อุดมศึกษา (สายวิชาการศึกษา)', 'อื่นๆ',
       'ไม่ทราบ', 'ต่ำกว่าประถม ศึกษา', 'ประถม ศึกษา'], dtype=object)
```

ไฟล์ที่ 2

```
  1 df_age_sex = pd.read_csv('/ผู้มีงานทำแยกตามเพศ,อายุ.csv')
  2 df_age_sex.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5832 entries, 0 to 5831
Data columns (total 8 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   year       5832 non-null   int64
 1   quarter    5832 non-null   object
 2   region     5832 non-null   object
 3   age_group  5832 non-null   object
 4   sex        5832 non-null   object
 5   value      5832 non-null   float64
 6   unit       5832 non-null   object
 7   source     5832 non-null   object
dtypes: float64(1), int64(1), object(6)
memory usage: 364.6+ KB
```

```
[7]  1 df_age_sex.isna().sum()
```

```
year         0
quarter      0
region       0
age_group    0
sex          0
value        0
unit         0
source       0
dtype: int64
```

```
[ ]  1 df_age_sex['age_group'].unique()
```

```
array(['รวม', '15-19 ปี', '20-24 ปี', '25-29 ปี', '30-34 ปี', '35-39 ปี',
       '40-49 ปี', '50-59 ปี', '60 ปีขึ้นไป'], dtype=object)
```

```
1 df_age_sex['region'].unique()
```

```
array(['ทั่วประเทศ ', 'กรุงเทพมหานคร', 'ภาคกลาง', 'ภาคเหนือ',
       'ภาคตะวันออกเฉียงเหนือ', 'ภาคใต้'], dtype=object)
```

Analysis

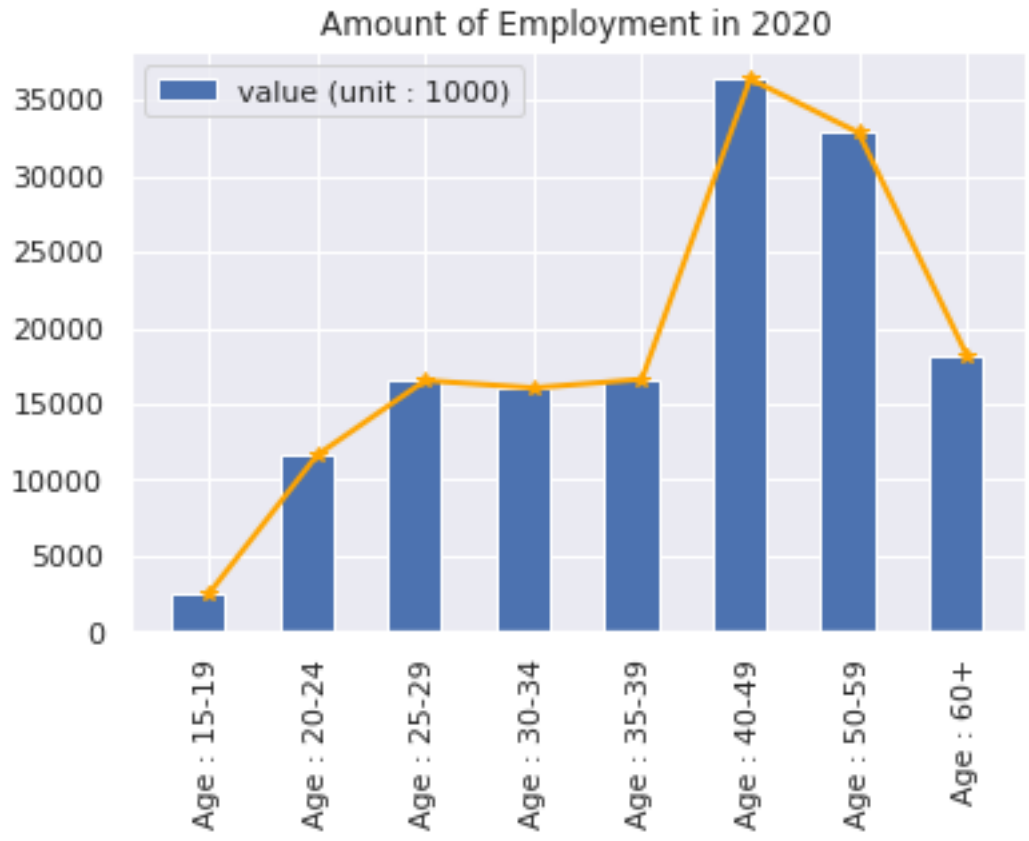
จากการทำการวิเคราะห์ข้อมูล การมีงานทำตามระดับการศึกษาที่สำเร็จ พบว่า ระดับการศึกษาที่สำเร็จมีมากที่สุดในประเทศไทย ตั้งแต่ปี พ.ศ. 2556-2563 คือ ระดับประถมศึกษา

```
## จำนวนคนทั้งหมดตามระดับการศึกษา ตั้งแต่ปี พ.ศ.2556-2563 ##
df_worker.loc[df_worker['level_of_edu'] == 'ต่ำกว่าประถม ศึกษา', 'level_of_edu'] = 'ต่ำกว่าประถมศึกษา'
df_worker.loc[df_worker['level_of_edu'] == 'ประถม ศึกษา', 'level_of_edu'] = 'ประถมศึกษา'
df_worker.loc[df_worker['level_of_edu'] == 'มัธยมศึกษาตอนปลาย (สายวิชาการศึกษา)', 'level_of_edu'] = 'มัธยมศึกษาตอนปลาย (สายสามัญ)'
df_worker.loc[df_worker['level_of_edu'] == 'อุดมศึกษา (สายวิชาการศึกษา)', 'level_of_edu'] = 'อุดมศึกษา (สายวิชาการ)'

group_level_edu = df_worker[(df_worker['region'] == 'ทั่วประเทศ') & (df_worker['area'] == 'รวม') & ~(df_worker['level_of_edu'] == 'รวม') & ~(df_worker['level_of_edu'] == 'ไม่ทราบ')][[
group_new_level_edu = group_level_edu.groupby(['level_of_edu'])
group_new_level_edu['value'].sum()
```

```
level_of_edu
ต่ำกว่าประถมศึกษา                  243987.03
ประถมศึกษา                          273617.40
มัธยมศึกษาตอนต้น                   199771.54
มัธยมศึกษาตอนปลาย (สายสามัญ)       153071.62
มัธยมศึกษาตอนปลาย (สายอาชีวศึกษา)   43188.29
อื่นๆ                                3670.47
อุดมศึกษา (สายวิชาการ)             185107.44
อุดมศึกษา (สายวิชาชีพ)              66837.08
ไม่มีการศึกษา                       38027.01
Name: value, dtype: float64
```

และ การมีงานทำจำแนกตามอายุและเพศ พบว่า ช่วงอายุ 40-49 ปี เป็นช่วงอายุที่มีงานทำมากที่สุดโดยรวมทั้งเพศชายและเพศหญิง เท่ากับ 36,367,000 คน ซึ่งตั้งแต่ปี พ.ศ.2556-2563 ช่วงอายุ 40-49 ปี ก็ยังเป็นช่วงอายุที่มีงานทำมากที่สุดเช่นกัน



ปัญหาที่พบ : การลงข้อมูลมีการลงข้อมูลเดียวกันแต่เป็นการพิมพ์คนละแบบ ทำให้ผลออกมามีจำนวนตัวแปรที่เยอะขึ้นแต่เป็นข้อมูลตัวเดียวกัน ทำให้ต้องมีการ replace ค่าให้ตรงกัน, ไม่ถนัดการเขียน Loop ทำให้ต้องมีการเขียนcodeซ้ำ, ทำอย่างไรให้ข้อมูลมีความน่าสนใจ, ข้อมูลมีการsumมาให้เป็นrowแล้ว ทำให้ต้องเขียนเงื่อนไขโดยที่ไม่ต้องรวม rowที่รวมมาให้ ทำให้ต้องเขียนcodeเงื่อนไขที่ยาวขึ้น