CREDIT ANALYSIS

0. Problem Solving

1. Python & MySQL

- Library
- Connect to MySQL
- Data info

2. Visualization & Analysis

- Customer age analysis chart: create a chart to understand how customer age is distributed in the data set.
- Customer gender division: determine the ratio of men and women in the data set.
- Annual income distribution (Income_Category) of customers.
- The chart can show the breakdown between card types (Card_Category) used by customers.
- Distribute customer's credit limit (Credit_Limit).
- Is there any relationship between customer age and age credit limit?
- Is there any relationship between education history and annual income of customers by age?
- Is there a correlation between the number of months of service and age limits?
- Is there any relationship between annual income and credit limit by age?

0. Problem Solving :

- **Customer age analysis chart: create a chart to understand how customer age is distributed in the data set.**
- **Customer gender division: determine the ratio of men and women in the data set.**
- **Annual income distribution (Income_Category) of customers.**
- **The chart can show the breakdown between card types (Card_Category) used by customers.**
- **Distribute customer's credit limit (Credit_Limit).**
- **Is there any relationship between customer age and age credit limit?**
- **Is there any relationship between education history and annual income of customers by age?**
- **Is there a correlation between the number of months of service and age limits?**
- **Is there any relationship between annual income and credit limit by age?**

1. Python & MySQL :

- **Library :**

In [ ]:
```
import seaborn as sns
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import mysql.connector
import random
sns.set_style("dark")
```

In [ ]:
```
def full_data():
    sql_query = """ SELECT * FROM credit """
    return sql_query
```

- Connect to MySQL :

In [ ]:
```
def get_database_from_MySQL_after_query(host, user, password, database, sql_quer
    # Thông tin kết nối
    config = {
        "host": host,    # Địa chỉ máy chủ MySQL
        "user": user,         # Tên người dùng MySQL
        "password": password, # Mật khẩu MySQL
        "database": database       # Tên cơ sở dữ liệu
    }

    # Kết nối tới MySQL
    conn = mysql.connector.connect(**config)

    # Tạo một đối tượng cursor
    cursor = conn.cursor()

    # Thực hiện truy vấn SQL
    cursor.execute(sql_query)

    # Trích xuất kết quả
    results = cursor.fetchall()

    # Trích xuất tên cột từ đối tượng cursor
    column_names = [i[0] for i in cursor.description]

    # Đóng kết nối
    conn.close()

    # Tạo DataFrame với tên cột
    df = pd.DataFrame(results, columns=column_names)

    return df
```

- Data info :

In [ ]:
```
df = get_database_from_MySQL_after_query("localhost","root","Khanhbg2522003","cr
df.head(10)
```

Out[ ]:

	CLIENT_ID	Customer_Age	Gender	Dependent_count	Education_Level	Marital_Status
0	768805383	45	M	3	High School	Married
1	818770008	49	F	5	Graduate	Single
2	713982108	51	M	3	Graduate	Married
3	769911858	40	F	4	High School	Unknown
4	709106358	40	M	3	Uneducated	Married
5	713061558	44	M	2	Graduate	Married
6	810347208	51	M	4	Unknown	Married
7	818906208	32	M	0	High School	Unknown
8	710930508	37	M	3	Uneducated	Single
9	719661558	48	M	2	Graduate	Single

In [ ]:
```
CLIENT_ID = "CLIENT_ID"
Customer_Age = "Customer_Age"
Gender = "Gender"
Dependent_count = "Dependent_count"
Education_Level = "Education_Level"
Marital_Status = "Marital_Status"
Income_Category = "Income_Category"
Card_Category = "Card_Category"
Months_on_book = "Months_on_book"
Credit_Limit  = "Credit_Limit"
```

```
*   CLIENT_ID: This can be a unique number representing each
customer or some type of personal identifier.
*   Customer_Age: Customer's age.
*   Gender: Customer's gender (M for male and F for female).
*   Dependent_count: Number of dependents of the customer.
*   Education_Level: Customer's education level.
*   Marital_Status: Customer's marital status.
*   Income_Category: Customer's annual income level.
*   Card_Category: The type of credit card used by the customer
(for example, Blue).
*   Months_on_book: Number of months that the customer has used
banking services.
*   Credit_Limit: Customer's credit limit.
```

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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10127 entries, 0 to 10126
Data columns (total 10 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   CLIENT_ID        10127 non-null  int64
 1   Customer_Age     10127 non-null  int64
 2   Gender           10127 non-null  object
 3   Dependent_count  10127 non-null  int64
 4   Education_Level  10127 non-null  object
 5   Marital_Status   10127 non-null  object
 6   Income_Category  10127 non-null  object
 7   Card_Category    10127 non-null  object
 8   Months_on_book   10127 non-null  int64
 9   Credit_Limit     10127 non-null  int64
dtypes: int64(5), object(5)
memory usage: 791.3+ KB
```

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

Out[ ]:

	CLIENT_ID	Customer_Age	Dependent_count	Months_on_book	Credit_Limit
count	1.012700e+04	10127.000000	10127.000000	10127.000000	10127.000000
mean	7.391776e+08	46.325960	2.346203	35.928409	8631.938679
std	3.690378e+07	8.016814	1.298908	7.986416	9088.788539
min	7.080821e+08	26.000000	0.000000	13.000000	1438.000000
25%	7.130368e+08	41.000000	1.000000	31.000000	2555.000000
50%	7.179264e+08	46.000000	2.000000	36.000000	4549.000000
75%	7.731435e+08	52.000000	3.000000	40.000000	11067.500000
max	8.283431e+08	73.000000	5.000000	56.000000	34516.000000

2. Visualzation :

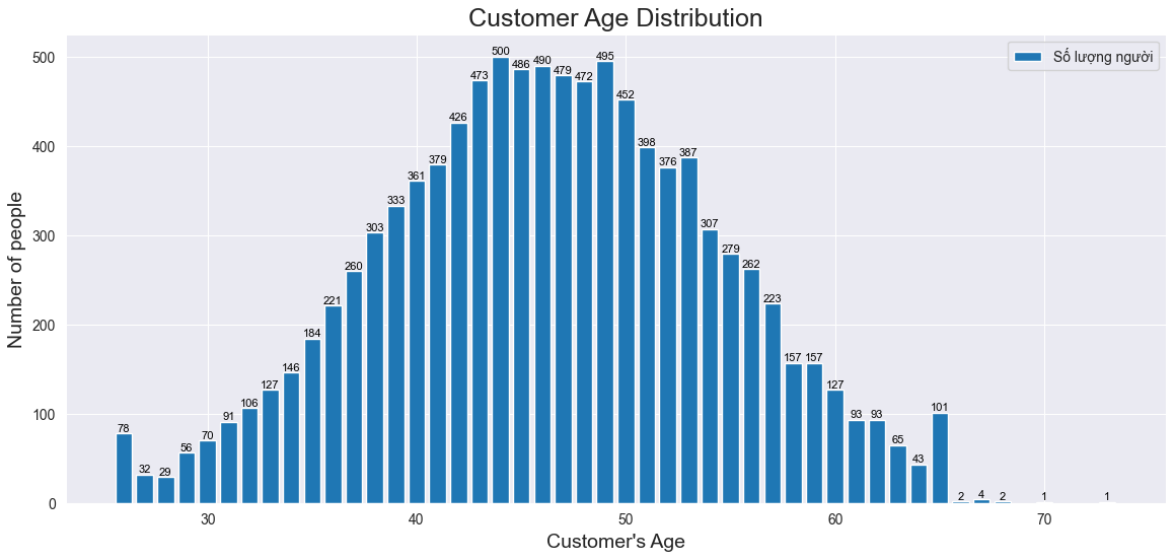
Customer age analysis chart: create a chart to understand how customer age is distributed in the data set. :

In [ ]:
```
df1 = df.groupby("Customer_Age")["CLIENT_ID"].count().sort_index().reset_index()
df1.head(5)
```

Out[ ]:

	Customer_Age	CLIENT_ID
0	26	78
1	27	32
2	28	29
3	29	56
4	30	70

In [ ]:
```
plt.figure(figsize=(14,6))
bars = plt.bar(df1["Customer_Age"], df1["CLIENT_ID"], label="Số lượng người")
plt.title("Customer Age Distribution", fontsize=18)
plt.xlabel("Customer's Age",fontsize=14)
plt.ylabel("Number of people",fontsize=14)
plt.grid()
plt.legend()
for bar in bars:
    yval = bar.get_height()
    plt.text(bar.get_x() + bar.get_width() / 2, yval, round(yval, 2), ha='center
```



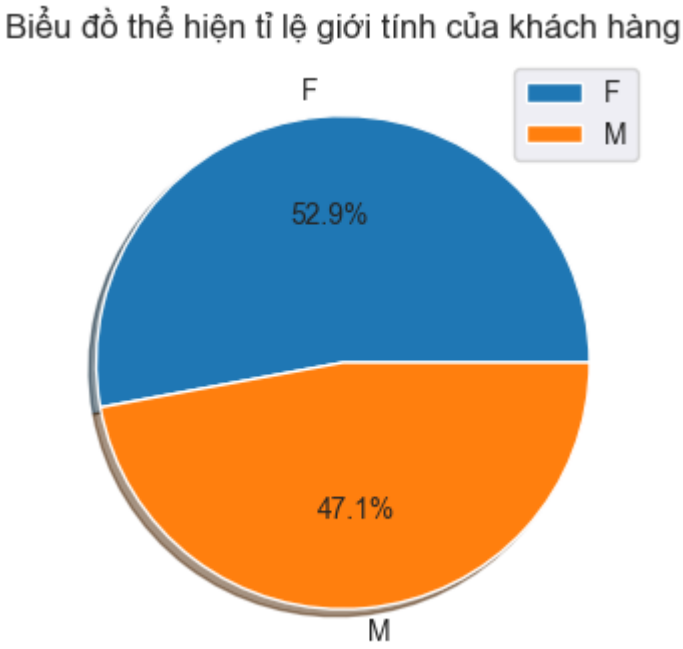
> Customers who use credit cards the most are between the ages of 40 and 50 years old

Customer gender division: determine the ratio of men and women in the data set.

In [ ]:
```
df2 = df.groupby("Gender")["CLIENT_ID"].count()
df2
```

Out[ ]:
```
Gender
F    5358
M    4769
Name: CLIENT_ID, dtype: int64
```

In [ ]:
```
plt.figure(figsize=(14,4))
plt.pie(df2.values,labels=df2.index,autopct='%1.1f%%',shadow=True)
plt.title("Biểu đồ thể hiện tỉ lệ giới tính của khách hàng")
plt.xticks()
plt.legend()
plt.show()
```



> The calculation limit rate for customers using the service is almost the same

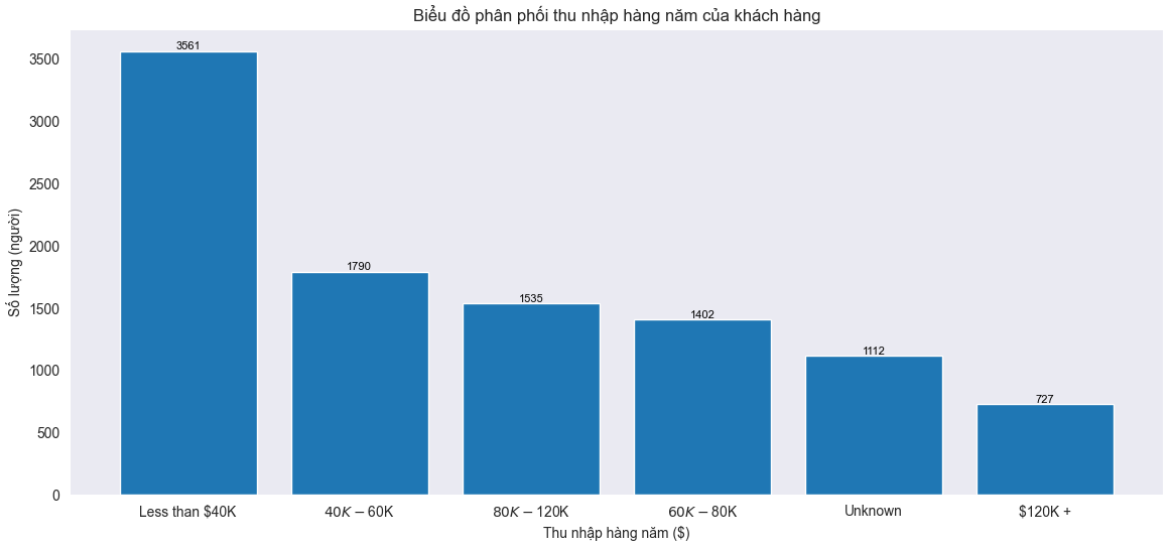
> The higher rate of women using credit cards may also mean they have more spending needs than men.

Annual income distribution (Income_Category) of customers.

In [ ]:
```
df5 = df.groupby(Income_Category)["CLIENT_ID"].count().sort_values(ascending=Fal
df5
```

Out[ ]:
```
Income_Category
Less than $40K    3561
$40K - $60K       1790
$80K - $120K      1535
$60K - $80K       1402
Unknown           1112
$120K +            727
Name: CLIENT_ID, dtype: int64
```

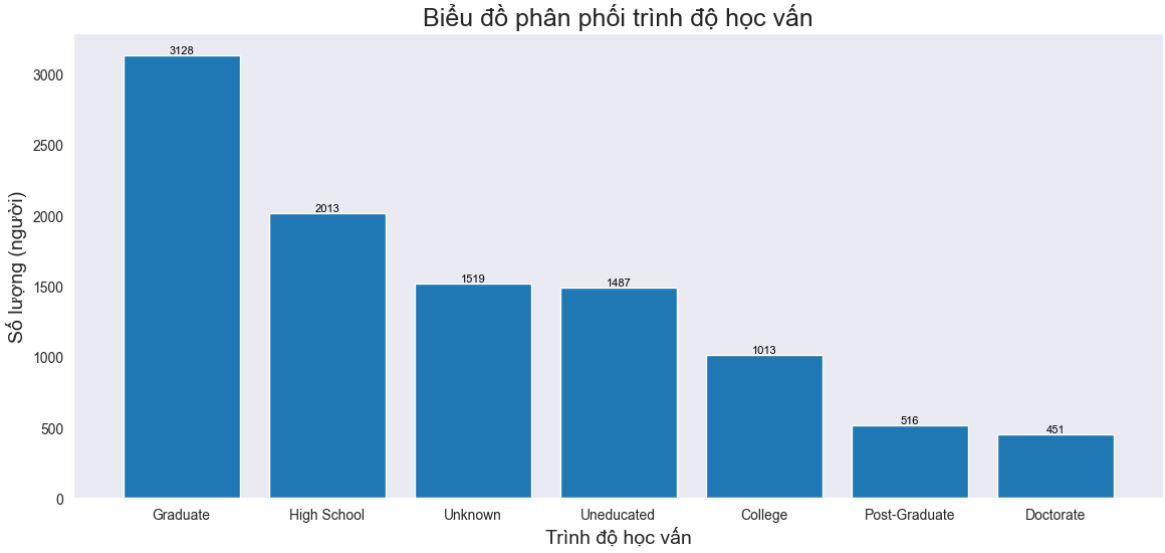
In [ ]:
```
plt.figure(figsize=(14,6))
bars = plt.bar(df5.index,df5.values)
plt.title("Biểu đồ phân phối thu nhập hàng năm của khách hàng")
plt.ylabel("Số lượng (người)")
plt.xlabel("Thu nhập hàng năm ($)")
for bar in bars:
    yval = bar.get_height()
    plt.text(bar.get_x() + bar.get_width() / 2, yval, round(yval, 2), ha='center
plt.show()
```



> Most credit card users have an average income of less than 40K$/year

In [ ]:
```
df3 = df.groupby(Education_Level)[CLIENT_ID].count().sort_values(ascending=False
```

In [ ]:
```
plt.figure(figsize=(14,6))
bars = plt.bar(df3.index,df3.values)
plt.title("Biểu đồ phân phối trình độ học vấn",fontsize=18)
plt.xlabel("Trình độ học vấn",fontsize=14)
plt.ylabel("Số lượng (người)",fontsize=14)
for bar in bars:
    yval = bar.get_height()
    plt.text(bar.get_x() + bar.get_width() / 2, yval, round(yval, 2), ha='center
plt.show()
```

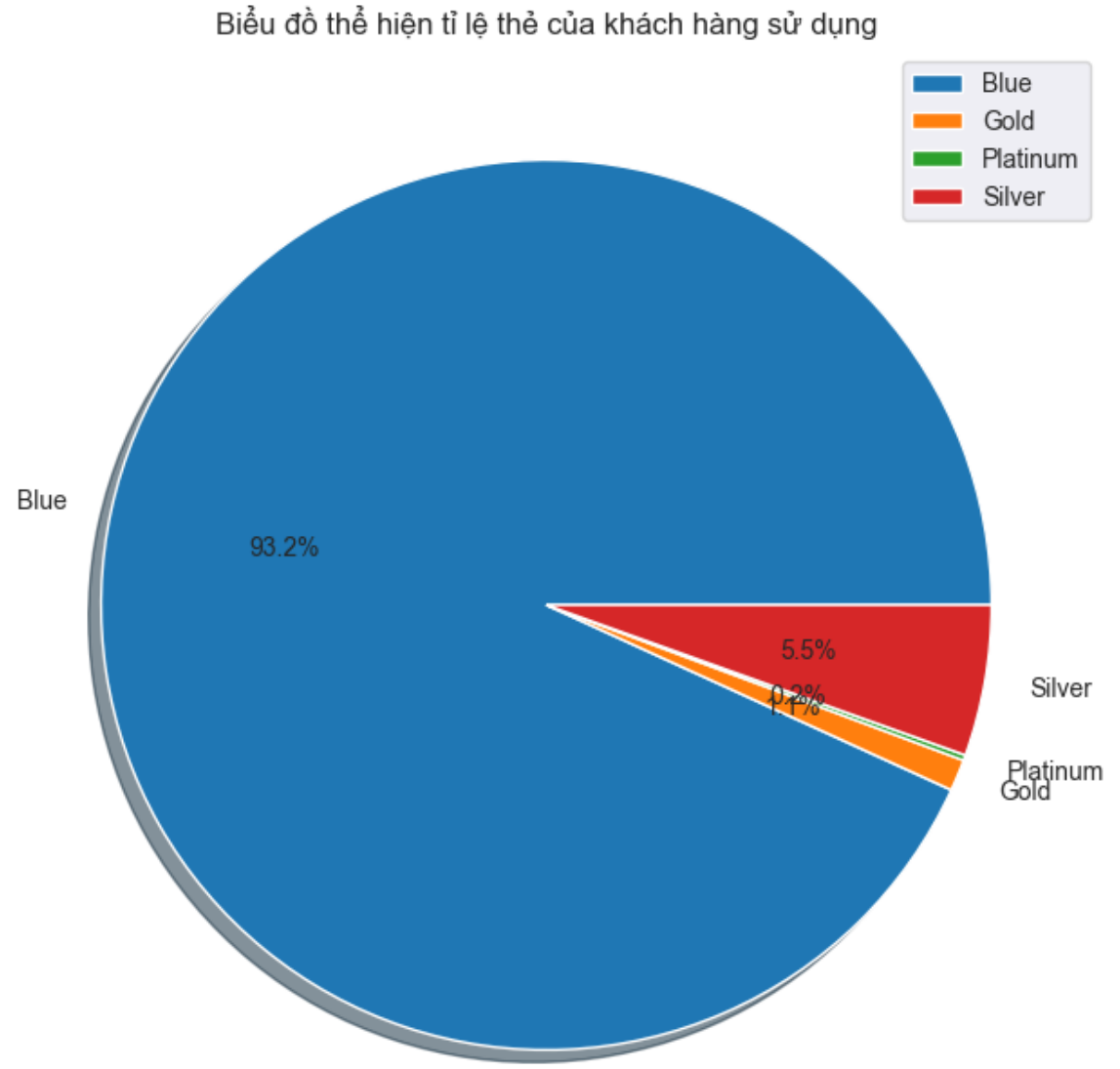


The chart can show the breakdown between card types (Card_Category) used by customers.

In [ ]:
```
df4 = df.groupby(Card_Category)["CLIENT_ID"].count()
df4.head(5)
```

Out[ ]:
```
Card_Category
Blue        9436
Gold         116
Platinum      20
Silver       555
Name: CLIENT_ID, dtype: int64
```

In [ ]:
```
plt.figure(figsize=(14,8))
plt.pie(df4.values,labels=df4.index,autopct='%1.1f%%',shadow=True)
plt.title("Biểu đồ thể hiện tỉ lệ thẻ của khách hàng sử dụng")
plt.xticks()
plt.legend()
plt.show()
```



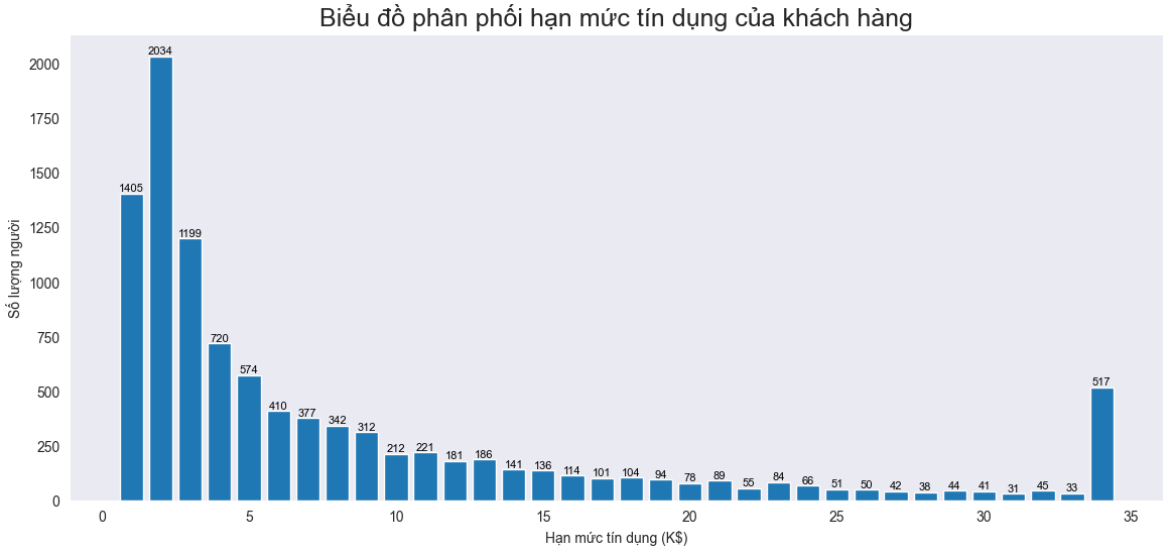
> The proportion of customers using green cards is the highest.

Distribute customer's credit limit (Credit_Limit).

In [ ]:
```
df6 = df.groupby(Credit_Limit)[CLIENT_ID].count().reset_index()
df6["Credit_Limit"] = (df6['Credit_Limit'] / 1000).astype(int)
```

```
df6 = df6.groupby(Credit_Limit)[CLIENT_ID].sum()
```

In [ ]:
```
plt.figure(figsize=(14,6))
bars = plt.bar(df6.index,df6.values)
plt.title("Biểu đồ phân phối hạn mức tín dụng của khách hàng",fontsize =18)
plt.xlabel("Hạn mức tín dụng (K$)")
plt.ylabel("Số lượng người")
for bar in bars:
    yval = bar.get_height()
    plt.text(bar.get_x() + bar.get_width() / 2, yval, round(yval, 2), ha='center
plt.show()
```

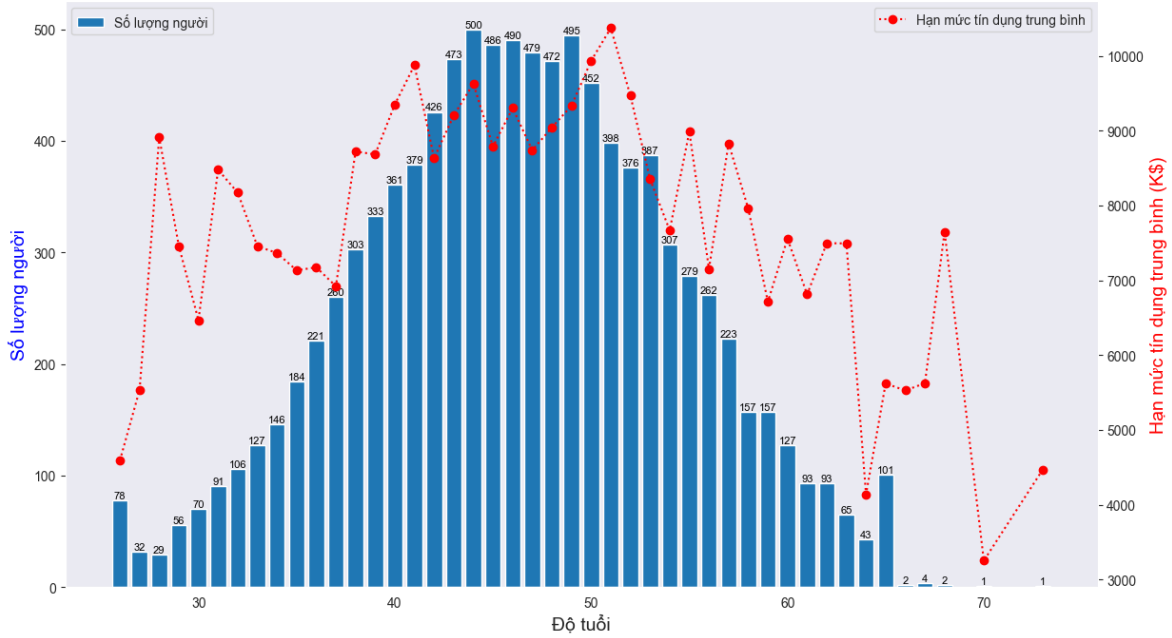


Is there any relationship between customer age and age credit limit?

In [ ]:
```
df7 = df.groupby(Customer_Age)[Credit_Limit].mean()
fig, ax1 = plt.subplots(figsize=(14,8 ))
bars = ax1.bar(df1["Customer_Age"], df1["CLIENT_ID"], label="Số lượng người")
ax1.set_title("Biểu đồ thể hiện sự tương quan giữa số lượng khách hàng và hạn mứ
ax1.set_xlabel("Độ tuổi",fontsize=14)
ax1.set_ylabel("Số lượng người", color="blue",fontsize=14)
plt.legend()
for bar in bars:
    yval = bar.get_height()
    plt.text(bar.get_x() + bar.get_width() / 2, yval, round(yval, 2), ha='center

ax2 = ax1.twinx()

# Vẽ biểu đồ hạn mức tín dụng
ax2.plot(df7.index, df7.values, color="red", label="Hạn mức tín dụng trung bình"
ax2.set_ylabel("Hạn mức tín dụng trung bình (K$)", color="red",fontsize=14)

# Hiển thị chú thích (Legend) cho cả hai trục
ax1.legend(loc='upper left')
ax2.legend(loc='upper right')
```

Out[ ]:
```
<matplotlib.legend.Legend at 0x23ce9465dd0>
```

Biểu đồ thể hiện sự tương quan giữa số lượng khách hàng và hạn mức tín dụng trung bình theo độ tuổi



> Although there are not many customers between the ages of 26 and 30, the average credit limit is very high, which is also evidence that young people have a need to spend a lot.

> Customers aged 40 to 50 are the largest and their average credit limit is also the highest.

> Customers aged 50 to 60 years old tend to decrease because their spending needs are low, so their average credit limit is also low.

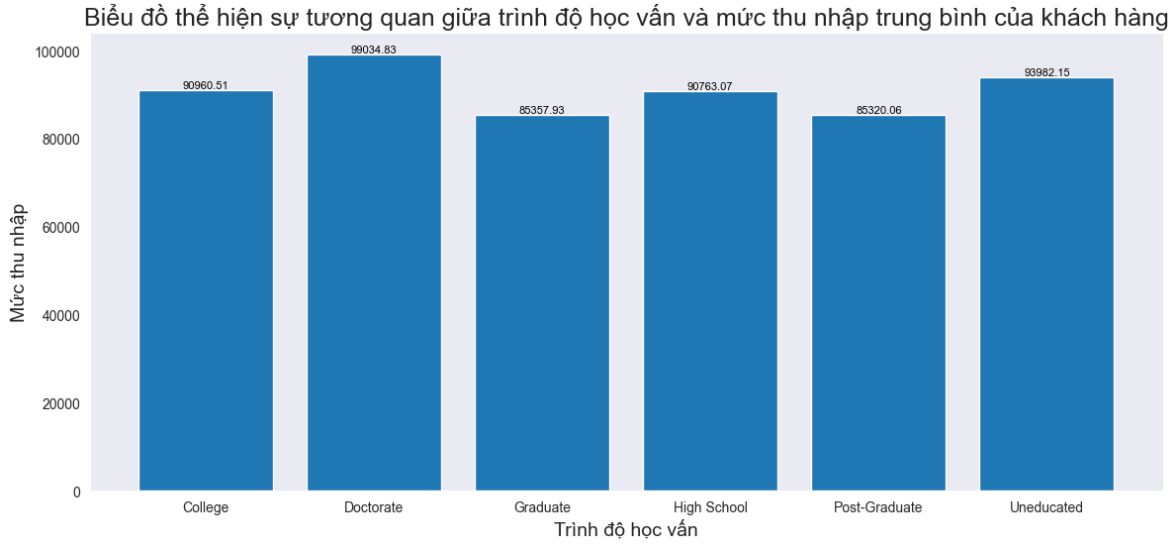
> In the US, people may go to a nursing home when they get old, so credit cards can be used during this time, so the average credit limit has also increased.

Is there any relationship between education history and annual income of customers by age?

In [ ]:
```
df3a = df.drop(df[(df["Income_Category"] == "Unknown") | (df["Education_Level"]
for i in df3a.index:
    if df3a["Income_Category"].loc[i] == "Less than $40K":
        df3a.at[i, "Income_Category"] = random.randint(0, 40000)
    elif df3a["Income_Category"].loc[i] == "$40K - $60K":
        df3a.at[i, "Income_Category"] = random.randint(40000, 60000)
    elif df3a["Income_Category"].loc[i] == "$60K - $80K":
        df3a.at[i, "Income_Category"] = random.randint(60000, 80000)
    elif df3a["Income_Category"].loc[i] == "$80K - $120K":
        df3a.at[i, "Income_Category"] = random.randint(80000, 120000)
    elif df3a["Income_Category"].loc[i] == "$120K +":
        df3a.at[i, "Income_Category"] = random.randint(120000, 1000000)
```

In [ ]:
```
df3b = df3a.groupby(Education_Level)[Income_Category].mean()
```

In [ ]:
```
plt.figure(figsize=(14,6))
bars = plt.bar(df3b.index,df3b.values)
plt.title("Biểu đồ thể hiện sự tương quan giữa trình độ học vấn và mức thu nhập
plt.xlabel("Trình độ học vấn",fontsize=14)
plt.ylabel("Mức thu nhập",fontsize=14)
for bar in bars:
    yval = bar.get_height()
    plt.text(bar.get_x() + bar.get_width() / 2, yval, round(yval, 2), ha='center
plt.show()
```



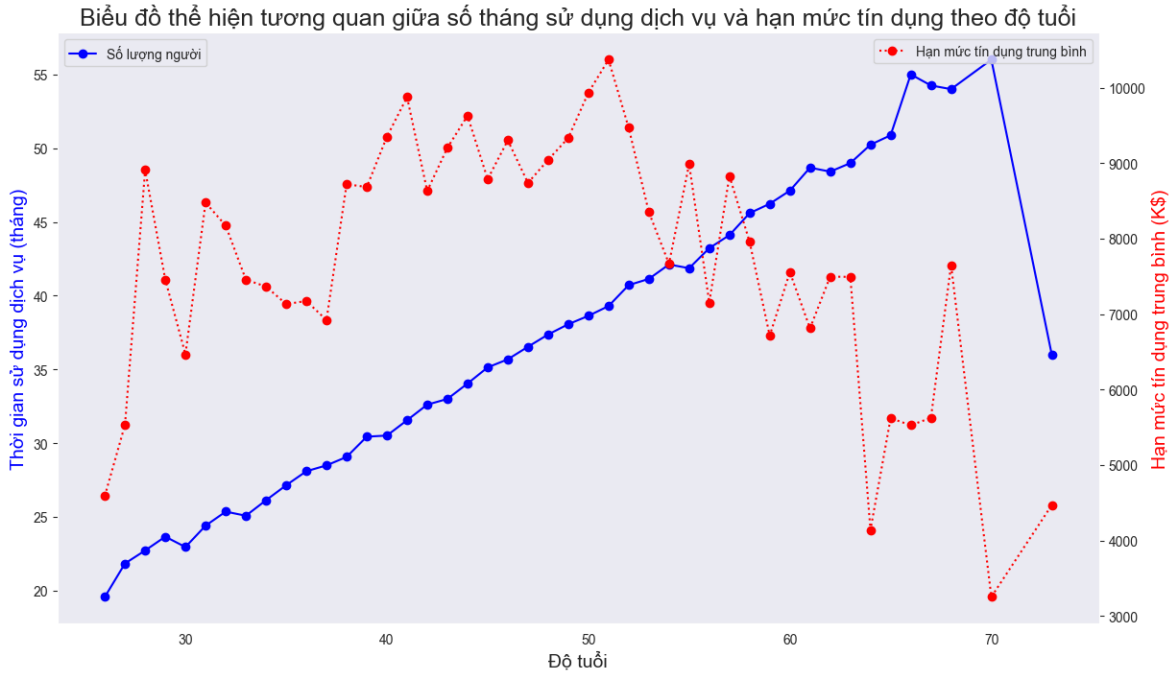
Is there a correlation between the number of months of service and age limits?

In [ ]:
```
df10 = df.groupby(Customer_Age)[Months_on_book].mean()
df10.head(5)
```

Out[ ]:
```
Customer_Age
26    19.576923
27    21.843750
28    22.724138
29    23.660714
30    22.957143
Name: Months_on_book, dtype: float64
```

In [ ]:
```
fig, ax1 = plt.subplots(figsize=(14,8 ))
ax1.plot(df10.index, df10.values, label="Số lượng người",color="blue",marker="o"
ax1.set_title("Biểu đồ thể hiện tương quan giữa số tháng sử dụng dịch vụ và hạn
ax1.set_xlabel("Độ tuổi",fontsize=14)
ax1.set_ylabel("Thời gian sử dụng dịch vụ (tháng)", color="blue",fontsize=14)
plt.legend()
ax2 = ax1.twinx()
# Vẽ biểu đồ hạn mức tín dụng
ax2.plot(df7.index, df7.values, color="red", label="Hạn mức tín dụng trung bình"
ax2.set_ylabel("Hạn mức tín dụng trung bình (K$)", color="red",fontsize=14)

# Hiển thị chú thích (legend) cho cả hai trục
ax1.legend(loc='upper left')
ax2.legend(loc='upper right')
```

Out[ ]: <matplotlib.legend.Legend at 0x23cebaca1d0>



Is there any relationship between annual income and credit limit by age?

- Hạn mức tin dụng trung bình của khách hàng với phân khúc từ 25 đến 30 tuổi khá cao mặc dù số lượng còn thấp.
- Hạn mức tín dụng trung bình của khách hàng với phân khúc từ 60 đến 70 tuổi
- Hạn mức tín dụng trung bình của khách hàng với phân khúc từ 30 đến 40 tuổi có xu hướng giảm dần vì trong khoảng thời gian đó họ đi làm với mức thu nhập cao nên có thể hạn chế việc sử dụng thẻ tín dụng điều này cũng có thể làm cho hạn mức tín dụng bị giảm xuống.

In [ ]:
```
df8 = df3a.groupby(Customer_Age)[Income_Category].mean()
df8.head(5)
```

Out[ ]:
```
Customer_Age
26          34933.9
27     71834.916667
28     22345.133333
29     53091.272727
30         40362.14
Name: Income_Category, dtype: object
```

In [ ]:
```
fig, ax1 = plt.subplots(figsize=(14,8 ))
ax1.plot(df8.index, df8.values, label="Số lượng người",color="blue",marker="o")
ax1.set_title("Biểu đồ thể hiện sự tương quan giữa thu nhập hàng năm và hạn mức
ax1.set_xlabel("Độ tuổi",fontsize=14)
ax1.set_ylabel("Thu nhập trung bình (K$)", color="blue",fontsize=14)
plt.legend()
ax2 = ax1.twinx()
# Vẽ biểu đồ hạn mức tín dụng
ax2.plot(df7.index, df7.values, color="red", label="Hạn mức tín dụng trung bình"
ax2.set_ylabel("Hạn mức tín dụng trung bình (K$)", color="red",fontsize=14)
```

```
# Hiển thị chú thích (legend) cho cả hai trục
ax1.legend(loc='upper left')
ax2.legend(loc='upper right')
```

Out[ ]: `<matplotlib.legend.Legend at 0x23cebb3be90>`

Biểu đồ thể hiện sự tương quan giữa thu nhập hàng năm và hạn mức tín dụng trung bình theo độ tuổi