In [37]:
```
import pandas as pd
from sklearn import preprocessing
import scipy as sc
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sea
print("Успішно")
```

```
Успішно
```

Переконуємось, що модулі успішно імпортовані

In [21]:
```
data = pd.read_csv("/Users/yurab/Desktop/Python for university/MMTAD/digi
#print(digital)
print(data.shape)
data_clean = data.dropna()
data_clean.to_csv("data_cleaned.csv", index=False)
print(data_clean.shape)
#data_wide = data_clean.pivot_table(index=[ "CampaignChannel"], columns="
#print(data_wide)
analize_column = ["Age", "Income", "AdSpend", "ClickThroughRate", "Conver
```

```
(8000, 20)
(8000, 20)
```

Як можна побачити, то датасет не містить жодних значень NA

Створюємо словар для збіру статистики

In [22]:
```
stats = {'Column': [], 'Mean': [], 'Trimmed': [], 'Median': [], 'Var': []
```

Шукаємо середнє значення по колонкам

In [23]:
```
for column in analize_column:
    mean = data_clean[column].mean()
    #print(mean)
    stats['Column'].append(column)
    stats['Mean'].append(mean)
    print(f"Статистика для {column}:")
    print(f"Середнє: {mean}")
    #print("\n")
#print(mean)
```

```
Статистика для Age:
Середнє: 43.6255
Статистика для Income:
Середнє: 84664.19675
Статистика для AdSpend:
Середнє: 5000.9448304235
Статистика для ClickThroughRate:
Середнє: 0.15482864919125
Статистика для ConversionRate:
Середнє: 0.104388735051875
Статистика для WebsiteVisits:
Середнє: 24.751625
Статистика для PagesPerVisit:
Середнє: 5.549299217260124
Статистика для TimeOnSite:
Середнє: 7.727718162508125
Статистика для SocialShares:
Середнє: 49.79975
Статистика для EmailOpens:
Середнє: 9.476875
Статистика для EmailClicks:
Середнє: 4.467375
Статистика для PreviousPurchases:
Середнє: 4.4855
Статистика для LoyaltyPoints:
Середнє: 2490.2685
Статистика для Conversion:
Середнє: 0.8765
```

Шукаємо зрізане середнє

In [24]:
```
for column in analize_column:
    trimmed = sc.stats.trim_mean(data_clean[column], proportiontocut=0.01
    stats['Trimmed'].append(trimmed)
    print(f"Статистика для {column}:")
    print(f"Зрізане середнє: {trimmed}")
```

```
Статистика для Age:
Зрізане середнє: 43.6280612244898
Статистика для Income:
Зрізане середнє: 84659.0663265306
Статистика для AdSpend:
Зрізане середнє: 4999.840146015944
Статистика для ClickThroughRate:
Зрізане середнє: 0.1548261275197704
Статистика для ConversionRate:
Зрізане середнє: 0.10437863952283165
Статистика для WebsiteVisits:
Зрізане середнє: 24.756760204081633
Статистика для PagesPerVisit:
Зрізане середнє: 5.550322727953445
Статистика для TimeOnSite:
Зрізане середнє: 7.727342948822959
Статистика для SocialShares:
Зрізане середнє: 49.80663265306122
Статистика для EmailOpens:
Зрізане середнє: 9.47640306122449
Статистика для EmailClicks:
Зрізане середнє: 4.466709183673469
Статистика для PreviousPurchases:
Зрізане середнє: 4.485204081632653
Статистика для LoyaltyPoints:
Зрізане середнє: 2490.094770408163
Статистика для Conversion:
Зрізане середнє: 0.8841836734693878
```

Шукаємо медіану для значень

In [25]:
```python
for column in analize_column:
    median = data_clean[column].median()
    stats['Median'].append(trimmed)
    print(f"Статистика для {column}:")
    print(f"Медіана: {median}")
```

```
Статистика для Age:
Медіана: 43.0
Статистика для Income:
Медіана: 84926.5
Статистика для AdSpend:
Медіана: 5013.440043500001
Статистика для ClickThroughRate:
Медіана: 0.1545054895
Статистика для ConversionRate:
Медіана: 0.104046464
Статистика для WebsiteVisits:
Медіана: 25.0
Статистика для PagesPerVisit:
Медіана: 5.534256644999999
Статистика для TimeOnSite:
Медіана: 7.682956376
Статистика для SocialShares:
Медіана: 50.0
Статистика для EmailOpens:
Медіана: 9.0
Статистика для EmailClicks:
Медіана: 4.0
Статистика для PreviousPurchases:
Медіана: 4.0
Статистика для LoyaltyPoints:
Медіана: 2497.0
Статистика для Conversion:
Медіана: 1.0
```

Шукаємо дисперсію по стовпчикам

In [26]:
```
for column in analize_column:
    var = data_clean[column].var()
    stats['Var'].append(var)
    print(f"Статистика для {column}:")
    print(f"Дисперсія: {var}")
```

```
Статистика для Age:
Дисперсія: 222.09301137642146
Статистика для Income:
Дисперсія: 1412285558.1228034
Статистика для AdSpend:
Дисперсія: 8054460.556349614
Статистика для ClickThroughRate:
Дисперсія: 0.007057210260389542
Статистика для ConversionRate:
Дисперсія: 0.003011628098131839
Статистика для WebsiteVisits:
Дисперсія: 204.8410399893734
Статистика для PagesPerVisit:
Дисперсія: 6.79831617671019
Статистика для TimeOnSite:
Дисперсія: 17.87782938744399
Статистика для SocialShares:
Дисперсія: 835.2773096011982
Статистика для EmailOpens:
Дисперсія: 32.61679233341701
Статистика для EmailClicks:
Дисперсія: 8.159955603825914
Статистика для PreviousPurchases:
Дисперсія: 8.34108238529822
Статистика для LoyaltyPoints:
Дисперсія: 2043547.906646084
Статистика для Conversion:
Дисперсія: 0.10826128266033716
```

Шукаємо стандартну дисперсію по стовпчиках

In [27]:
```
for column in analize_column:
    std = data_clean[column].std()
    stats['Std'].append(std)
    print(f"Статистика для {column}:")
    print(f"Дисперсія: {std}")
```

```
Статистика для Age:
Дисперсія: 14.902785356315828
Статистика для Income:
Дисперсія: 37580.387945347284
Статистика для AdSpend:
Дисперсія: 2838.0381527297363
Статистика для ClickThroughRate:
Дисперсія: 0.08400720362200817
Статистика для ConversionRate:
Дисперсія: 0.05487830261708027
Статистика для WebsiteVisits:
Дисперсія: 14.312268862391225
Статистика для PagesPerVisit:
Дисперсія: 2.607358083714278
Статистика для TimeOnSite:
Дисперсія: 4.228218228455574
Статистика для SocialShares:
Дисперсія: 28.901164502510937
Статистика для EmailOpens:
Дисперсія: 5.711111304590116
Статистика для EmailClicks:
Дисперсія: 2.8565636005217727
Статистика для PreviousPurchases:
Дисперсія: 2.8880932092469283
Статистика для LoyaltyPoints:
Дисперсія: 1429.5271619126668
Статистика для Conversion:
Дисперсія: 0.32903082326787736
```

Шукаємо медіану по стовпчиках

In [28]:
```
for column in analize_column:
    mad = data_clean[column].median()
    stats['Mad'].append(mad)
    print(f"Статистика для {column}:")
    print(f"Медіана: {mad}")
```

```
Статистика для Age:
Медіана: 43.0
Статистика для Income:
Медіана: 84926.5
Статистика для AdSpend:
Медіана: 5013.440043500001
Статистика для ClickThroughRate:
Медіана: 0.1545054895
Статистика для ConversionRate:
Медіана: 0.104046464
Статистика для WebsiteVisits:
Медіана: 25.0
Статистика для PagesPerVisit:
Медіана: 5.534256644999999
Статистика для TimeOnSite:
Медіана: 7.682956376
Статистика для SocialShares:
Медіана: 50.0
Статистика для EmailOpens:
Медіана: 9.0
Статистика для EmailClicks:
Медіана: 4.0
Статистика для PreviousPurchases:
Медіана: 4.0
Статистика для LoyaltyPoints:
Медіана: 2497.0
Статистика для Conversion:
Медіана: 1.0
```

In [29]:
```python
for column in analize_column:
    a_mad = sc.stats.median_abs_deviation(data_clean[column])
    stats['A_mad'].append(a_mad)
    print(f"Статистика для {column}:")
    print(f"Абсол. мед. відхилення: {a_mad}")
```

```
Статистика для Age:
Абсол. мед. відхилення: 13.0
Статистика для Income:
Абсол. мед. відхилення: 32454.0
Статистика для AdSpend:
Абсол. мед. відхилення: 2449.0594789999996
Статистика для ClickThroughRate:
Абсол. мед. відхилення: 0.07274559850000001
Статистика для ConversionRate:
Абсол. мед. відхилення: 0.047838236000000006
Статистика для WebsiteVisits:
Абсол. мед. відхилення: 12.0
Статистика для PagesPerVisit:
Абсол. мед. відхилення: 2.275300944500001
Статистика для TimeOnSite:
Абсол. мед. відхилення: 3.7002447840000006
Статистика для SocialShares:
Абсол. мед. відхилення: 25.0
Статистика для EmailOpens:
Абсол. мед. відхилення: 5.0
Статистика для EmailClicks:
Абсол. мед. відхилення: 2.0
Статистика для PreviousPurchases:
Абсол. мед. відхилення: 3.0
Статистика для LoyaltyPoints:
Абсол. мед. відхилення: 1222.0
Статистика для Conversion:
Абсол. мед. відхилення: 0.0
```

Імпортуємо статистичний збір

In [35]:
```
statistics_df = pd.DataFrame(stats)
statistics_df.to_csv('statistics.csv', index=False)
```

Нормалізуємо набір даних

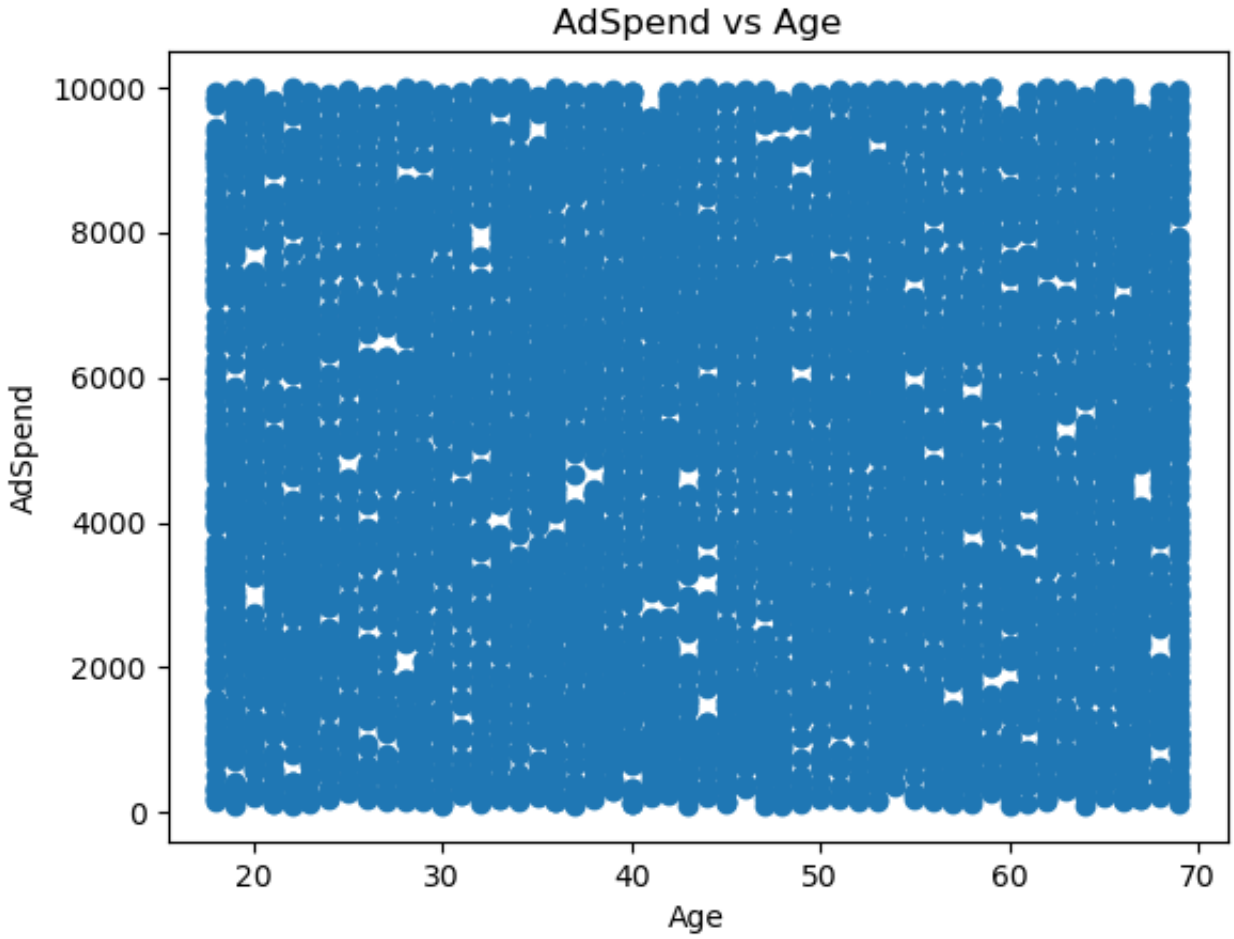
In [30]:
```
d = preprocessing.normalize(data_clean[["Age", "Income", "AdSpend", "Clic
scaled_df = pd.DataFrame(d, columns=["Age", "Income", "AdSpend", "ClickTh
#scaled_df.head(10)
scaled_df
scaled_df.to_csv("scaled_df.csv", index=False)
```

Нормалізуємо мін.мах

In [31]:
```
scaler = preprocessing.MinMaxScaler(feature_range=(0, 2))
c = scaler.fit_transform(data_clean[["Age", "Income", "AdSpend", "ClickTh
scaled_df2 = pd.DataFrame(c, columns=["Age", "Income", "AdSpend", "ClickT
#scaled_df2.head()
scaled_df2
scaled_df2.to_csv("scaled_df2.csv", index=False)
```

Формування графіка 1

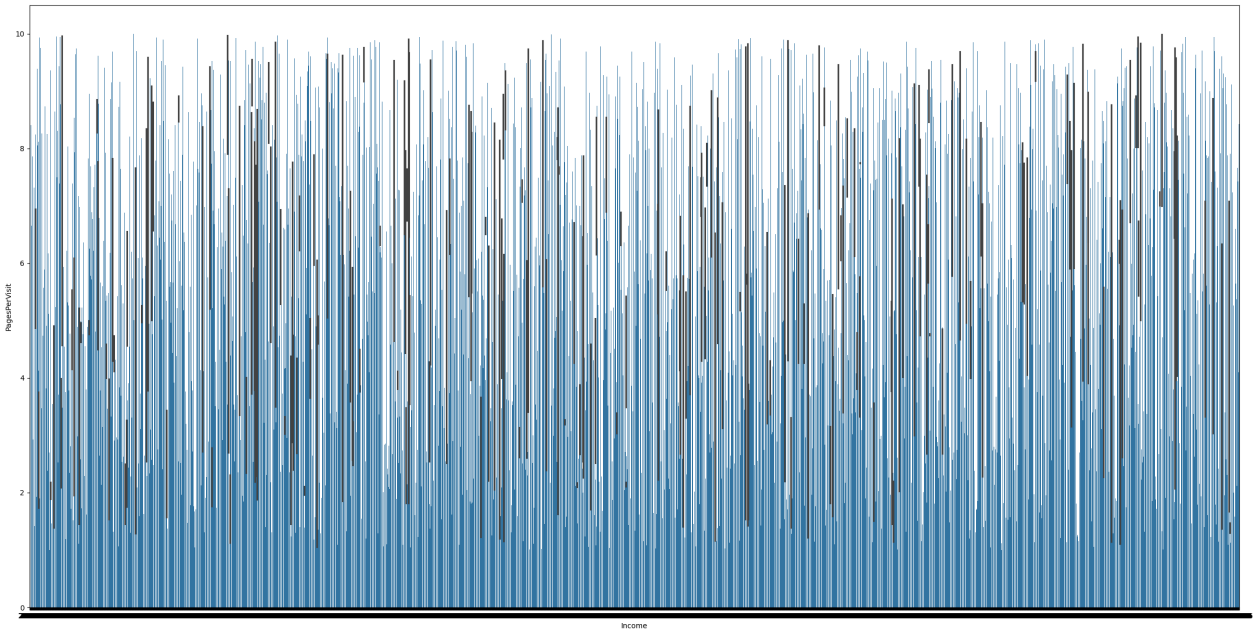
In [32]:
```
plt.figure()
plt.scatter(data_clean["Age"], data_clean["AdSpend"])
plt.xlabel('Age')
plt.ylabel('AdSpend')
plt.title('AdSpend vs Age')
plt.show()
```



Формування графіка 2

In [33]:
```
plt.figure(figsize=(30, 15))
sea.barplot(x = data_clean["Income"], y = data_clean["PagesPerVisit"])
```

Out[33]:
```
<Axes: xlabel='Income', ylabel='PagesPerVisit'>
```



Формування графіка 3

In [34]:
```
plt.figure(figsize=(30, 15))
sea.barplot(x = data_clean["Age"], y = data_clean["TimeOnSite"])
```

Out[34]: <Axes: xlabel='Age', ylabel='TimeOnSite'>