Super WholeSale Store Customer Database Analysis

A super wholesale store is collecting data based off their customer's membership card.The store sells a wide variety of goods such as perishable foods, electronics, clothing and furniture.

Through collecting this data, we want to help find the answers to the following questions:

1) What is the occupation distribution of shop's ideal customers?

2) Customers of which Profession spends the most amount at the shop?

3) How is Family Size associated with the Annual Income?

4) Is there a correlation between customer annual income and the Spending Score?

5) Is there a relationship between Spending Score and Family Size distributed across Gender?

6)Is there a certain popular Profession amongst each Gender?

7) Is there a relationship between Age and Spending Score across Age?

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

In [5]:
```
df=pd.read_csv('Customers.csv')
df.head()
```

Out[5]:

	CustomerID	Gender	Age	Annual Income ($)	Spending Score (1-100)	Profession	Work Experience	Family Size
0	1	Male	19	15000	39	Healthcare	1	4
1	2	Male	21	35000	81	Engineer	3	3
2	3	Female	20	86000	6	Engineer	1	1
3	4	Female	23	59000	77	Lawyer	0	2
4	5	Female	31	38000	40	Entertainment	2	6

The Customer Dataset Attributes:

- Customer ID
- Gender
- Age
- Annual Income ($)
- Spending Score

> Score assigned by the shop, based on customer behavior and spending nature

- Profession
- Work Experience (Years)
- Family Size

In [ ]:
```
df['Gender'].value_counts()
```

Out[ ]:
```
Female    1186
Male       814
Name: Gender, dtype: int64
```

Research Question Visualizations and Analyses:

1. What is the occupation of the shop's ideal customers?

In [ ]:
```
df['Profession'].value_counts()
```

Out[ ]:
```
Artist           612
Healthcare       339
Entertainment    234
Engineer         179
Doctor           161
Executive        153
Lawyer           142
Marketing         85
Homemaker         60
Name: Profession, dtype: int64
```

In [ ]:
```
names = df['Profession'].value_counts().index.tolist()
names
```

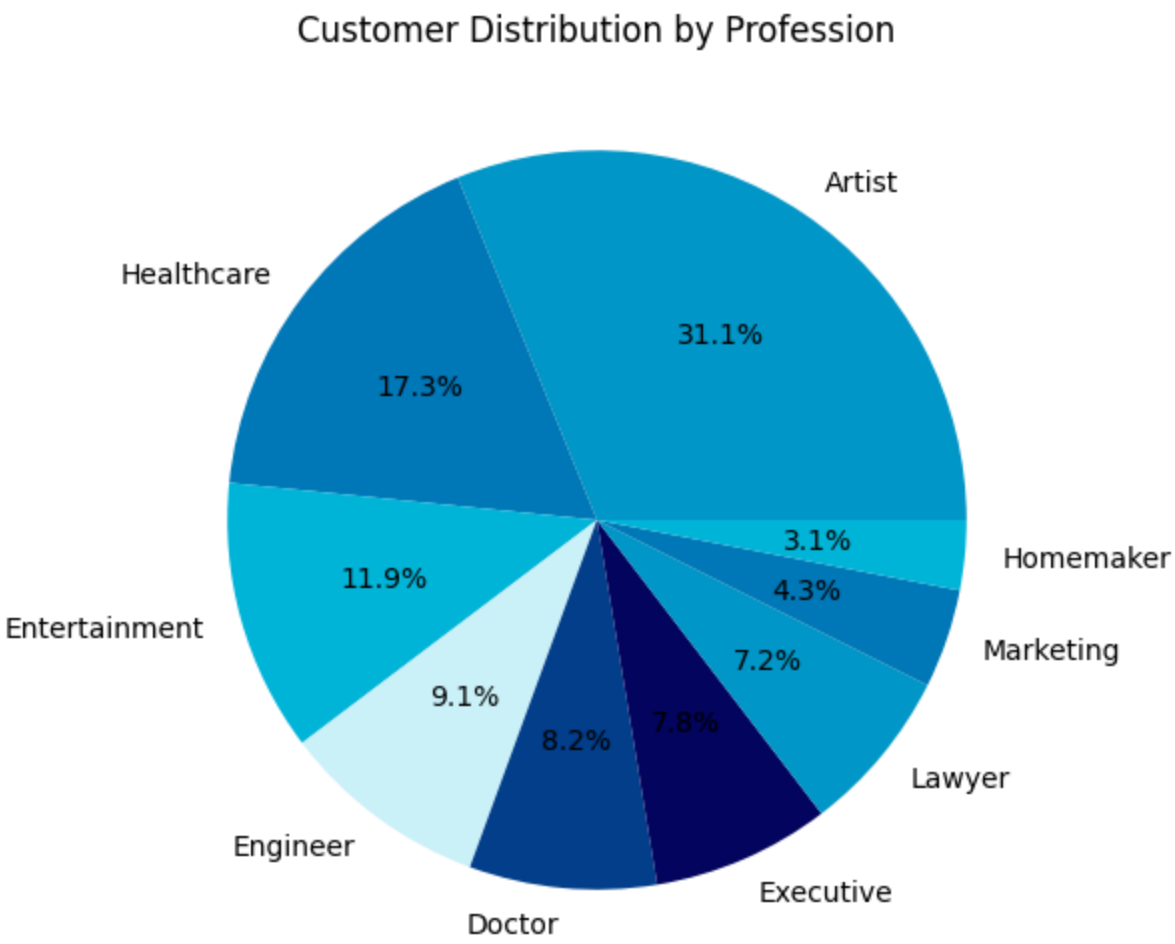
Out[ ]:
```
['Artist',
 'Healthcare',
 'Entertainment',
 'Engineer',
 'Doctor',
 'Executive',
 'Lawyer',
 'Marketing',
 'Homemaker']
```

In [ ]:
```
values = df['Profession'].value_counts().tolist()
values
```

Out[ ]:
```
[612, 339, 234, 179, 161, 153, 142, 85, 60]
```

In [ ]:
```
color_parental = ['#0096c7', '#0077b6', '#00b4d8', '#caf0f8', '#023e8a', '#030
```

In [ ]:
```
fig = plt.figure(figsize=(6, 6))
plt.pie(values, labels=names, colors = color_parental, autopct='%1.1f%%')
plt.title('Customer Distribution by Profession')
plt.show()
```

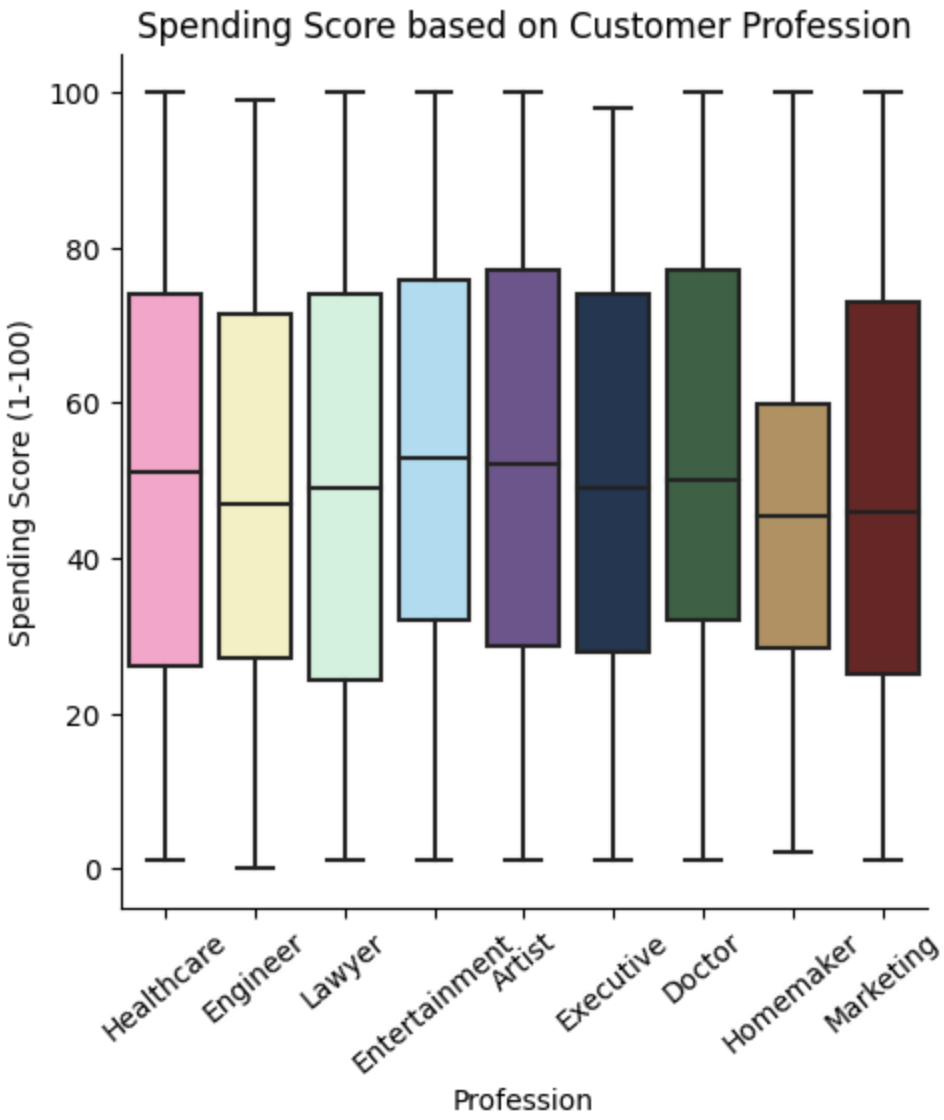


Interpretation:

The highest number of the shops's ideal customers are Art professionals with 31.1% where as the lowest amount of their ideal customers are Homemakers consisting about 3.1%.

2. Which Profession has the highest spending at the store?

In [ ]:
```
sns.catplot(data=df, x='Profession', y='Spending Score (1-100)', kind='box',
     palette = {'Healthcare':'#ff99c8', 'Engineer':'#fcf6bd', 'Lawyer':'#d0f4d
            'Entertainment':'#a9def9', 'Artist':'#6a4c93', 'Executive':'#1d355
            'Doctor':'#386641', 'Homemaker':'#bb9457', 'Marketing':'#6f1d1b' }
plt.title('Spending Score based on Customer Profession')
plt.xticks(rotation = 40)
plt.show()
```



Interpretation:

All professions have a median Spending score between 40 and 60.

Customers who are in the Entertainment profession have the highest median Spending Score at approximately 55.

Customers who are Artists have the highest upper quartile Spending score at nearly 80.

3. How is Family Size associated with Annual Income?

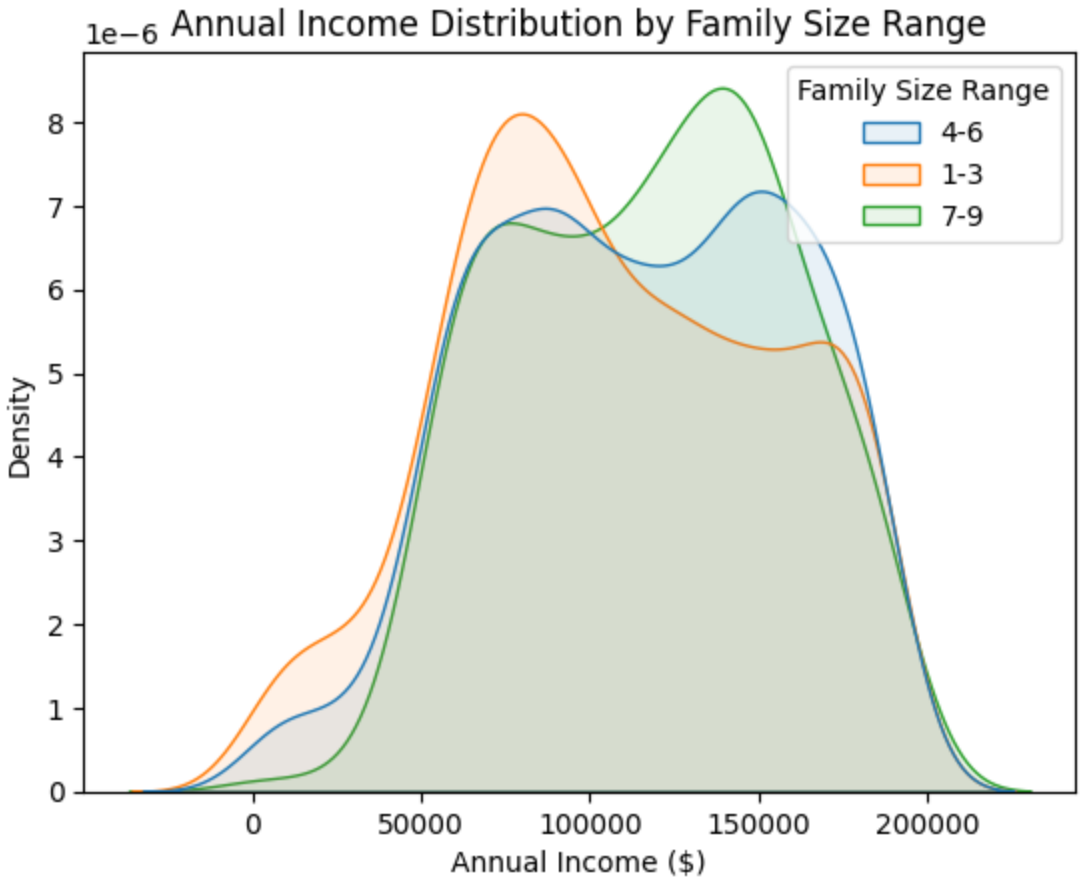
In [8]:
```
def famsize_bucket(x):
  if x >= 7:
    return '7-9'
  elif x >= 4:
    return '4-6'
  else:
    return '1-3'
```

In [9]:
```
df['Family Size Range'] = df['Family Size'].apply(famsize_bucket)
df.head(3)
```

Out[9]:

	CustomerID	Gender	Age	Annual Income ($)	Spending Score (1-100)	Profession	Work Experience	Family Size	Family Size Range
0	1	Male	19	15000	39	Healthcare	1	4	4-6
1	2	Male	21	35000	81	Engineer	3	3	1-3
2	3	Female	20	86000	6	Engineer	1	1	1-3

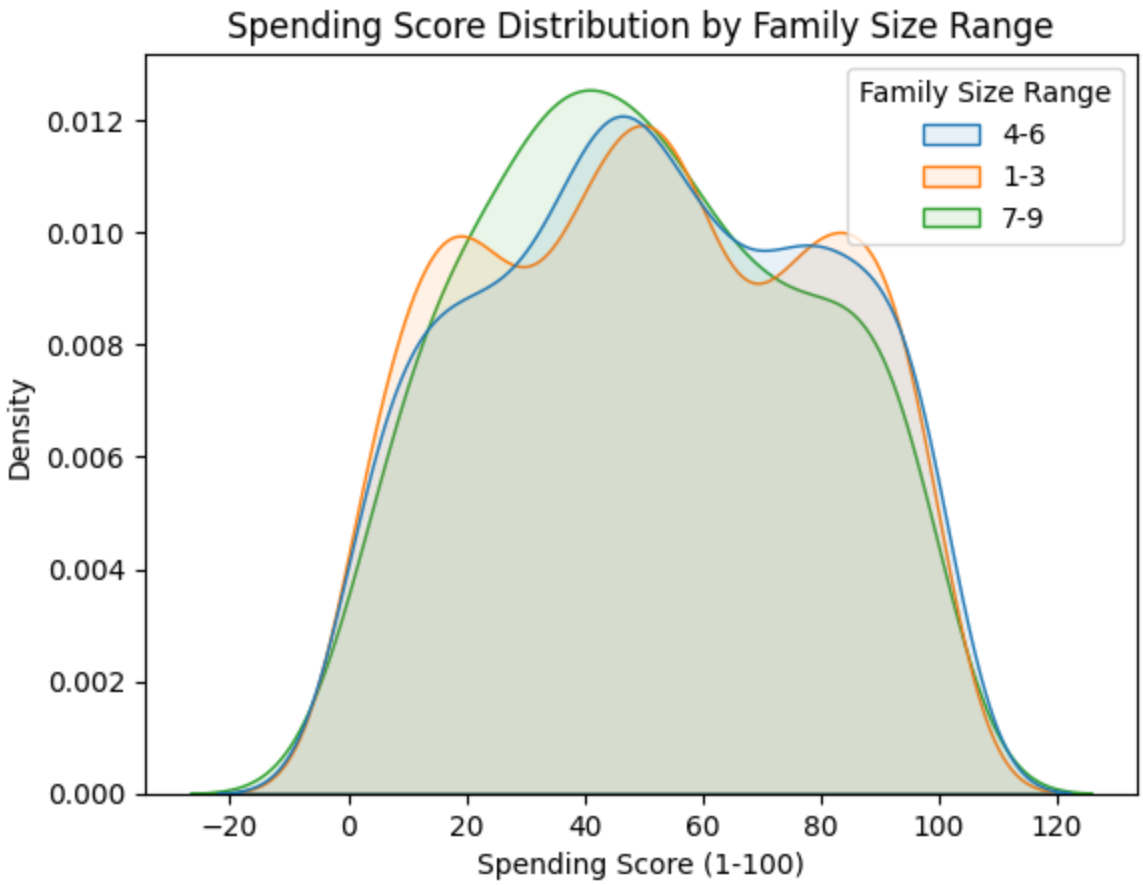
In [ ]:
```
sns.kdeplot(data=df, x='Annual Income ($)', hue='Family Size Range', fill=True
            alpha=0.1, common_norm=False)
plt.title('Annual Income Distribution by Family Size Range')
plt.show()
```



3b. How is Family Size associated with Spending Score?

The above visualization prompted us to dig further to analyze if there was any insights we could gain by comparing Family Size to Spending Score:

In [ ]:
```
sns.kdeplot(data=df, x='Spending Score (1-100)', hue='Family Size Range', fill=
            alpha=0.1, common_norm=False)
plt.title('Spending Score Distribution by Family Size Range')
plt.show()
```



Interpretation:

Annual Income Distrubution for customers with family sizes that range from 1-3 typically have a median income lower than $100,000.

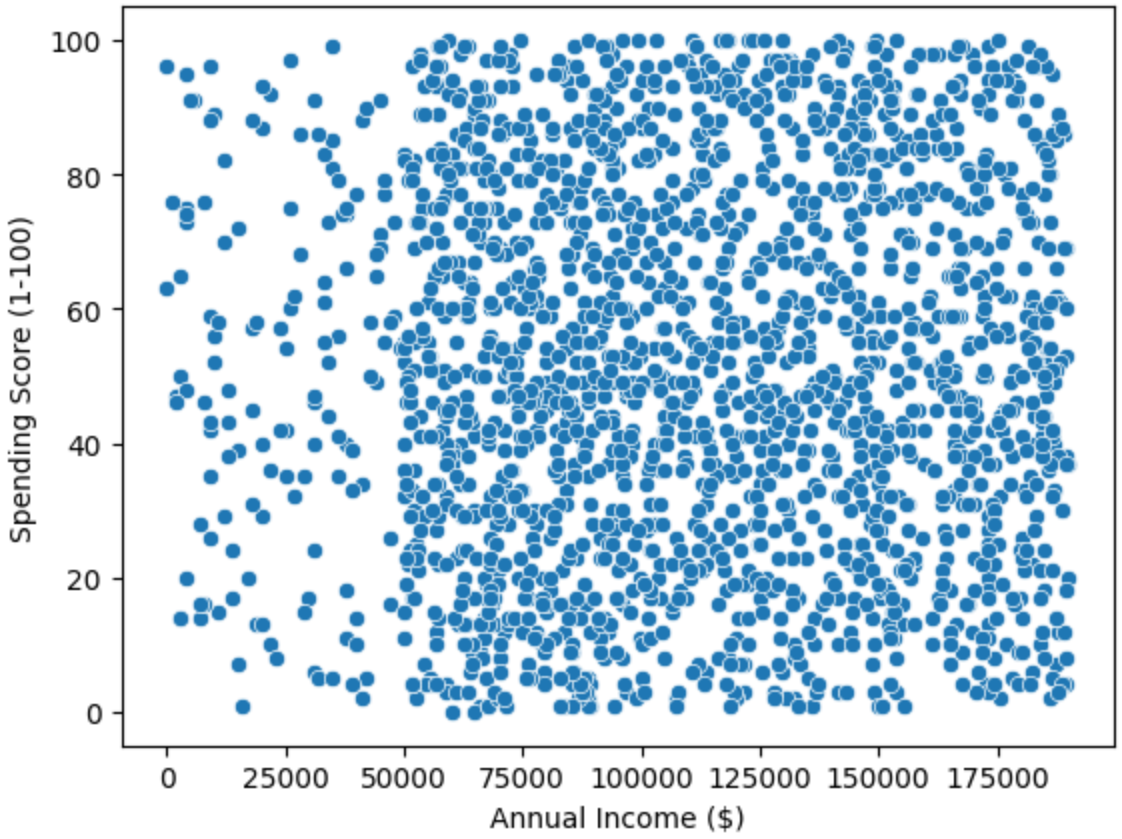
Alternatively, customers with family sizes that range from 7-9 have the highest median income at nearly $125,000.

Although customers with family sizes ranging from 7-9 have the highest median income, their spending score is leaning toward the lower side compared to customers with smaller families.

This could be a potential indication that customers with larger families are less likely to purchase goods at the super store due to affordability.

4. Is there a correlation between customer annual income and the store's Spending Score?

In [ ]:
```
sns.scatterplot(data=df, x='Annual Income ($)', y='Spending Score (1-100)')
plt.show()
```



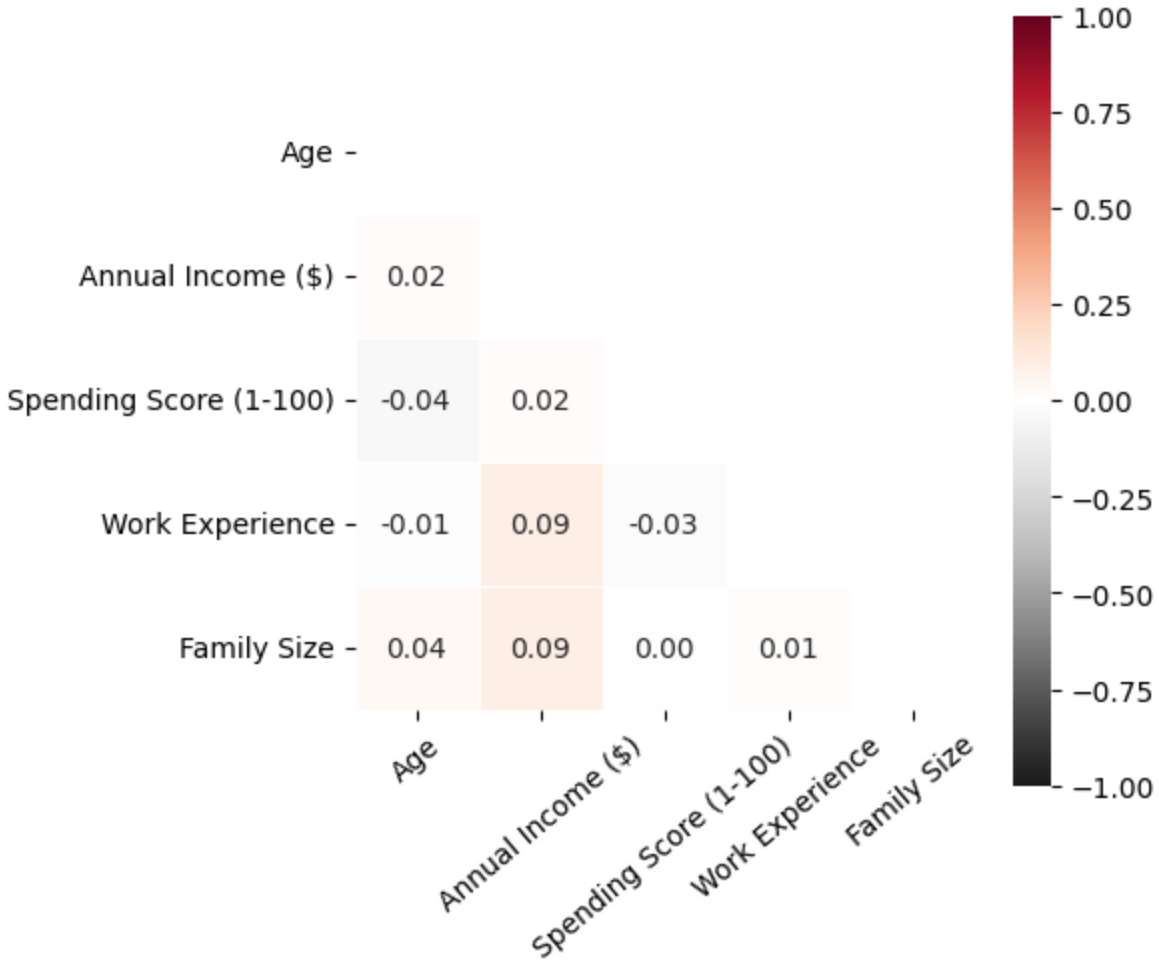
In [ ]:
```
new_df = df.iloc[:, 1:9]
coco = new_df.corr()
coco
```

```
<ipython-input-67-8d56dac65d31>:2: FutureWarning: The default value of numeric
_only in DataFrame.corr is deprecated. In a future version, it will default to
False. Select only valid columns or specify the value of numeric_only to silen
ce this warning.
  coco = new_df.corr()
```

Out[ ]:

	Age	Annual Income ($)	Spending Score (1-100)	Work Experience	Family Size
Age	1.000000	0.021378	-0.041798	-0.014319	0.038254
Annual Income ($)	0.021378	1.000000	0.023299	0.089136	0.093005
Spending Score (1-100)	-0.041798	0.023299	1.000000	-0.028948	0.002232
Work Experience	-0.014319	0.089136	-0.028948	1.000000	0.011873
Family Size	0.038254	0.093005	0.002232	0.011873	1.000000

In [ ]:
```
tri_matrix = np.triu(coco)
plt.figure(figsize = (5, 5))
sns.heatmap(coco, square=True, cmap='RdGy_r',fmt='.2f', annot=True, linewidth=(
            mask=tri_matrix)
plt.xticks(rotation=40)
plt.show()
```



Interpretation:

We can see through the above scatterplot and correlation coefficient heatmap that there is not a strong correlation between customer Annual Income and their Spending Score.

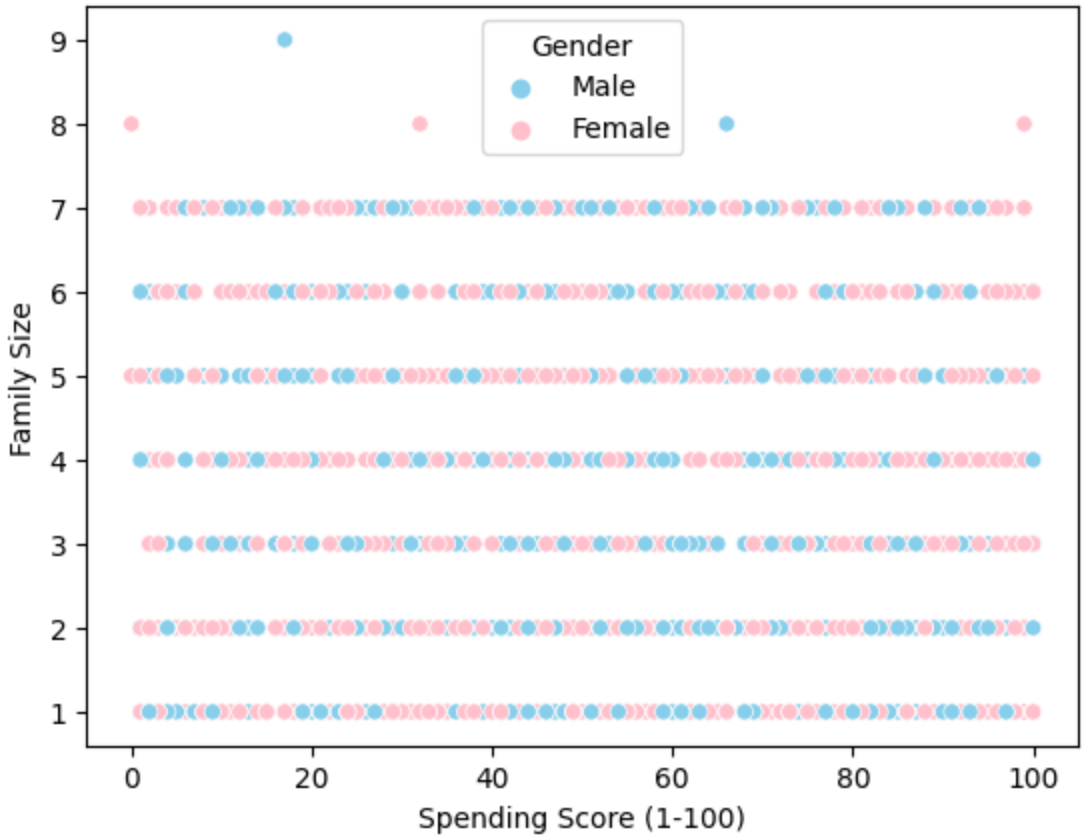
The scatterplot does show us that most customers' annual income is above $50,000. However there is no relation of this variable to spending score.

The heatmap tells us that correlation between Annual Income and Spending Score is 0.02. This is the second lowest correlation coefficient in the heatmap.

*Note: It appears that all correlation coefficients between variables are quite low with the highest values being 0.09 for Work Experience vs. Annual Income and Family Size vs. Annual Income.

5. Is there a relationship between Spending Score and Family Size distributed across Gender?

In [ ]:
```
sns.scatterplot(data=df,x='Spending Score (1-100)',y='Family Size',hue='Gender
plt.show()
```

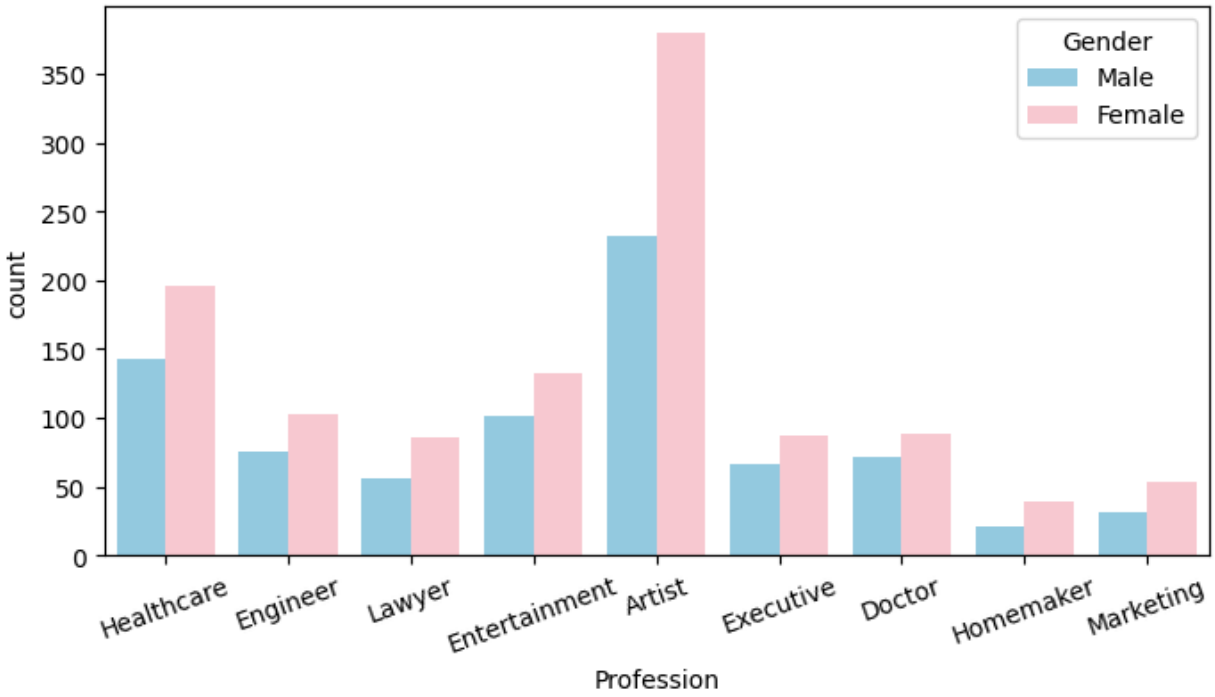


Interpretation:

As the graph shows, there seems to be no relationship between Spending Score and Family whether the buyer is Male or Female.

6. Is there a certain popular Profession amongst each Gender?

In [ ]:
```
plt.figure(figsize=(8,4))
sns.countplot(data=df,x='Profession', hue='Gender',
              palette={'Female':'pink', 'Male':'skyblue'})
plt.xticks(rotation=20)
plt.show()
```

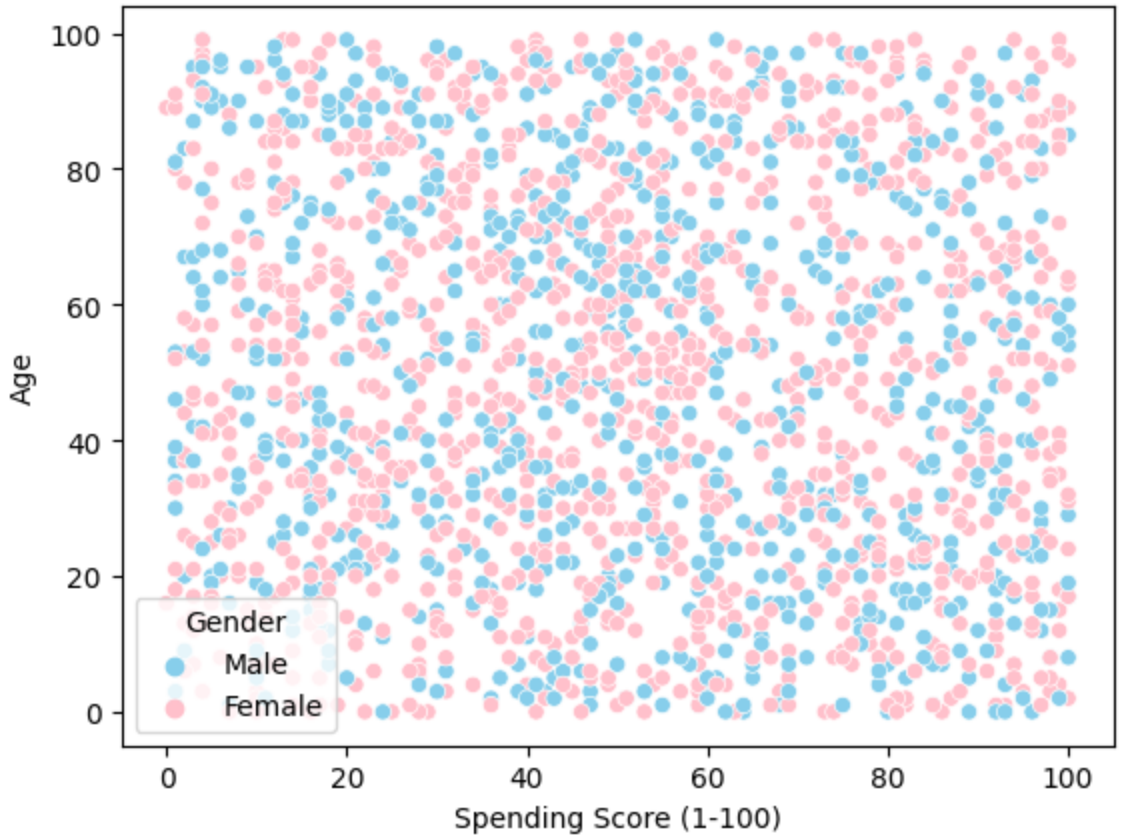


Interpretation:

As shown in the pie chart previously, the highest Profession is in the Art industry. However, the bar graph also shows how in each profession category, Female is the highest in each profession.

7. Is there a relationship between Age and Spending Score?

In [ ]:
```
sns.scatterplot(data=df,x='Spending Score (1-100)',y='Age',hue='Gender',palett
plt.show()
```



Interpretation:

There is not a clear correlation between Gender, Age, and Spending Score.

Conclusion:

In analyzing the customer database attributes, we were able to identify the following key points:

- More than half of the store's customers are categorized as having Artist (31%), Healthcare (17%) and Entertainment (12%) Professions.
    - There are majority women within each Profession.
- Customers in Artist and Entertainment Professions have higher median to upper quartile Spending Scores than most other professions.
- Customers with a family size ranging from 7-9 tend to have a **higher** annual income but a **lower** spending score than customers with smaller families.
- There is no clear correlation between Age and Spending Score.

The store could assess the prices for each product type that customers tend to purchase to see how the attributes in the dataset can apply. For example, customers with larger family sizes might have a lower spending score due to potential unaffordability of necessary products.

The store could also evaluate their product offerings to target their customers' key attributes, some of which include those in an Artist or Entertainment profession or those that are female.

Through the analysis above, our team was able to assist the Super Wholesale Store in determining some key attributes about their customer database that could inform future business decisions and strategies.

In [2]:
```
!jupyter nbconvert --to html group_project_\(Customers\).ipynb
```

```
[NbConvertApp] Converting notebook group_project_(Customers).ipynb to html
[NbConvertApp] Writing 2256919 bytes to group_project_(Customers).html
```