CSCE 5310: Methods in Empirical Analysis

Project Proposal
Analysis on Store sales and Customer Data
Spring 2023

Group Members

Student ID	Name	Group
11546725	Pragnesh Kumar Devarakonda	Group2
11580431	Viswak sena palaparthi	Group2
11648583	Pavan Kalyan Kumar boddu	Group2
11609230	Madhuri Sri Yarramareddy	Group2

Table of Contents

Summary Of Research

Introduction

This document is used to explain the analysis of the dataset which contains the data related to the Customer details who have membership in a shop. The data explains the details of customers with few fields like profession, Annual income.

The dataset which will help the shop owner to understand about the business in the store.

Motivation and Background

The dataset includes 2000 records and 8 columns which will explain about the customers who have membership with the store. The details include Customer ID, Gender, Age, Annual income, Spending Score-Score assigned by the shop, based on customer behavior and spending nature, Profession, Work experience, Family Size.

With this data the shop owner will have complete details about his business. There comes the analysis of the purchases and profits in the business. The dataset will play a key role in understanding the business, which is very important for the store owner, includes increasing the stocks based on the spending score of the customers, selling different types of items based on the customers need, etc.

Dataset and details

The Dataset includes the details of customers of a store.


Customers.csv

The research includes data of customer as the sample data to analysis the business of the store.
We got this dataset from the URL below.
https://www.kaggle.com/datasets/datascientistanna/customers-dataset?resource=download

Exploratory Data Analysis:

We performed an insights analysis on the customer data obtained from Kaggle by plotting and graphing the data. Based on the information provided in the data, we drew conclusions for the questions outlined in our proposal. The analysis was conducted using Python code to visualize the data.

- Python libraries such as NumPy, Pandas, Matplotlib, Plotly, and Seaborn were used for plotting and analysis of the customer data.
- Read csv. File using pandas libraries and imported that to a Dataframe df.
- Customer datasets have a total of 2000 records with fields Customer ID, Gender, Age, Annual Income ($), Spending Score (1-100), Profession, Work Experience, Family Size fields.
- Python library Pandas will help us to read the dataset csv file into a data frame.
- Below is the dataframe,

Using pandas libraries read the customer.csv file into a dataframe df, below is the dataframe.

	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
...	...	...	...	...	...	...	...	...
1995	1996	Female	71	184387	40	Artist	8	7
1996	1997	Female	91	73158	32	Doctor	7	7
1997	1998	Male	87	90961	14	Healthcare	9	2
1998	1999	Male	77	182109	4	Executive	7	2
1999	2000	Male	90	110610	52	Entertainment	5	2

2000 rows × 8 columns

Found 35 records are having Null values in professions, below are the null value records from the dataframe.

	CustomerID	Gender	Age	Annual Income ($)	Spending Score (1-100)	Profession	Work Experience	Family Size
79	80	Female	49	98000	42	NaN	1	1
118	119	Female	51	84000	43	NaN	2	7
219	220	Female	59	76000	61	NaN	9	1
237	238	Male	95	36000	35	NaN	0	4
437	438	Male	76	136259	14	NaN	0	7
440	441	Female	0	57373	29	NaN	0	7
498	499	Male	95	121725	3	NaN	12	3
545	546	Female	89	107359	26	NaN	10	6
601	602	Male	61	126370	20	NaN	11	4
641	642	Male	66	121377	19	NaN	7	7
665	666	Male	28	101414	64	NaN	8	1
703	704	Male	22	114011	40	NaN	5	7
801	802	Male	81	148208	36	NaN	5	7
817	818	Female	91	154456	71	NaN	0	7
850	851	Male	69	186655	32	NaN	7	2
903	904	Female	15	174501	37	NaN	9	7
927	928	Male	25	81367	87	NaN	0	3
1009	1010	Male	69	61637	67	NaN	0	5
1067	1068	Female	30	78821	46	NaN	1	4
1088	1089	Female	10	162076	78	NaN	0	3

We tried to fill the Null values based on percentage of each profession and percentage of gender in the dataset with the missing records in the professions.

Below is the code,

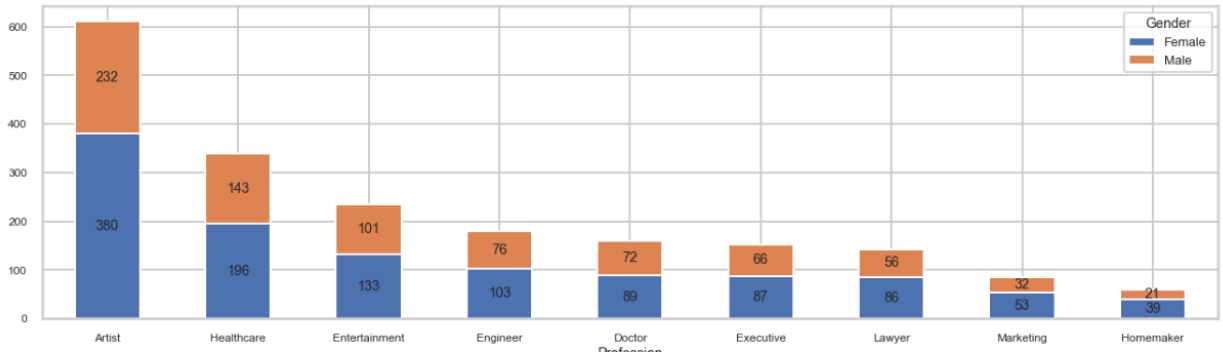
```
# Calculate the percentage of each Profession for available data
profession_pct = df[~df['Profession'].isna()].groupby(['Profession'])['Profession'].count() / len(df[~df['Profession'].isna()])

# Calculate the percentage of each gender for available data
gender_pct = df.groupby(['Gender'])['Gender'].count() / len(df)

# Fill in missing profession values
for i, row in df[df['Profession'].isna()].iterrows():
    profession = np.random.choice(profession_pct.index, p=profession_pct.values)
    gender = row['Gender']
    df.at[i, 'Profession'] = profession
    df.at[i, 'Gender'] = gender
```

The First line of code calculates the percentage of each profession in the available data. It filters out any rows where the profession value is missing, then groups the remaining rows by profession and counts the number of rows in each group. The count for each profession is then divided by the total number of rows where profession value is not missing to get percentage.

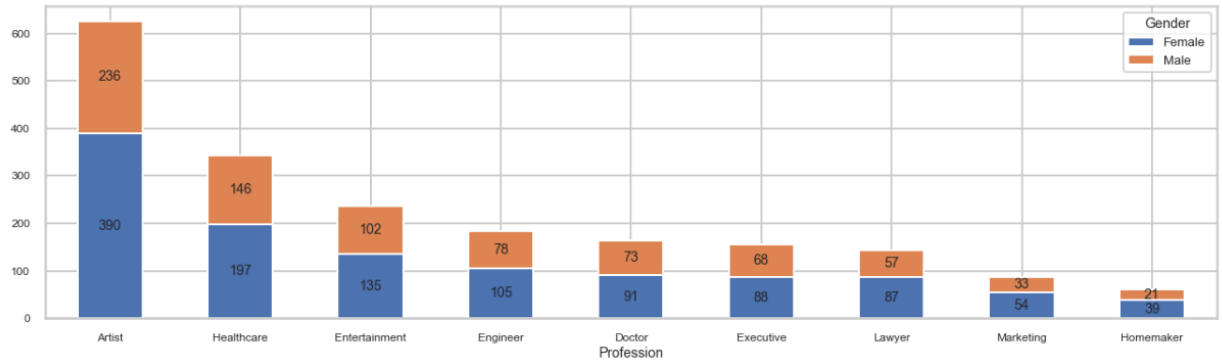
The Second line of code calculates the percentage of each gender in the entire dataset by grouping the rows by gender and counting the number of rows in each group, then dividing the count for each gender by the total number of rows in the dataset.

The last block of code fills in missing profession values by iterating over each row in the dataset where the profession value is missing. For each row, it randomly selects a profession from the available professions based on the percentage of each profession calculated in the first line of code. It also assigns the gender values from the row to the ‘Gender’ column and updates the ‘profession’ column with the randomly selected profession.

Before the filling, the count of Male and female in each profession are below.



After filling the Null values below are the count of Male, Female with respect to the Professions.

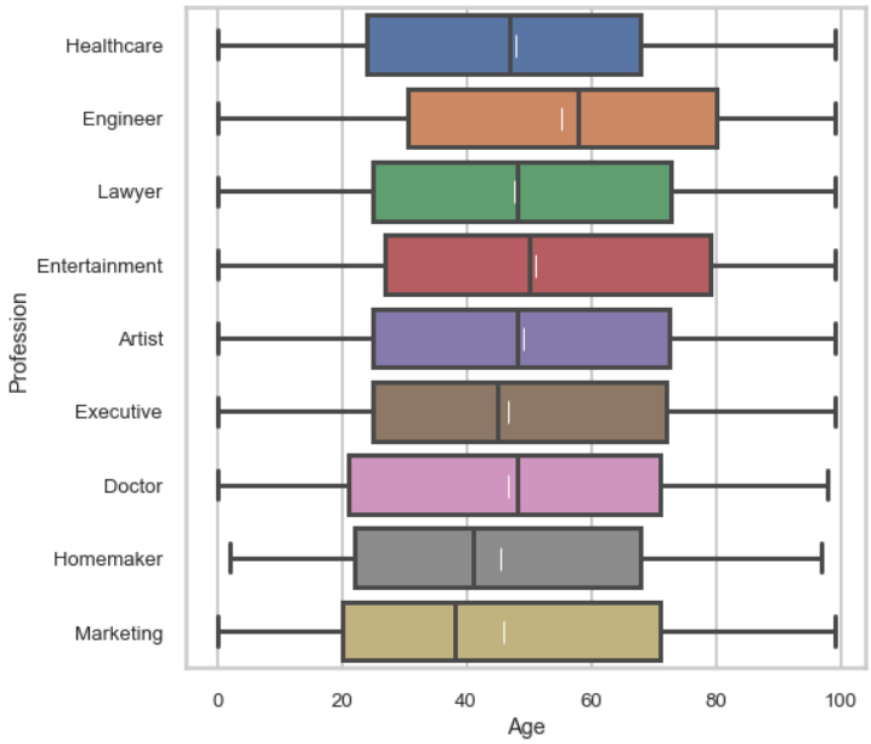


Prediction Questions:

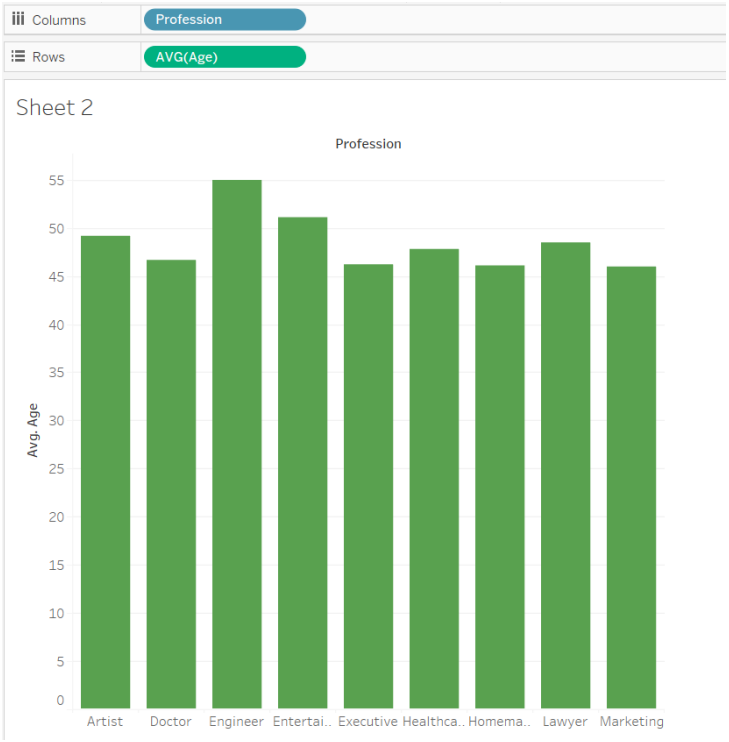
Can we predict which profession customers with an age above 25 more likely to have a membership.

Tried to insight the data using pandas, Seaborn libraries. The graph shows Age verse Professions.

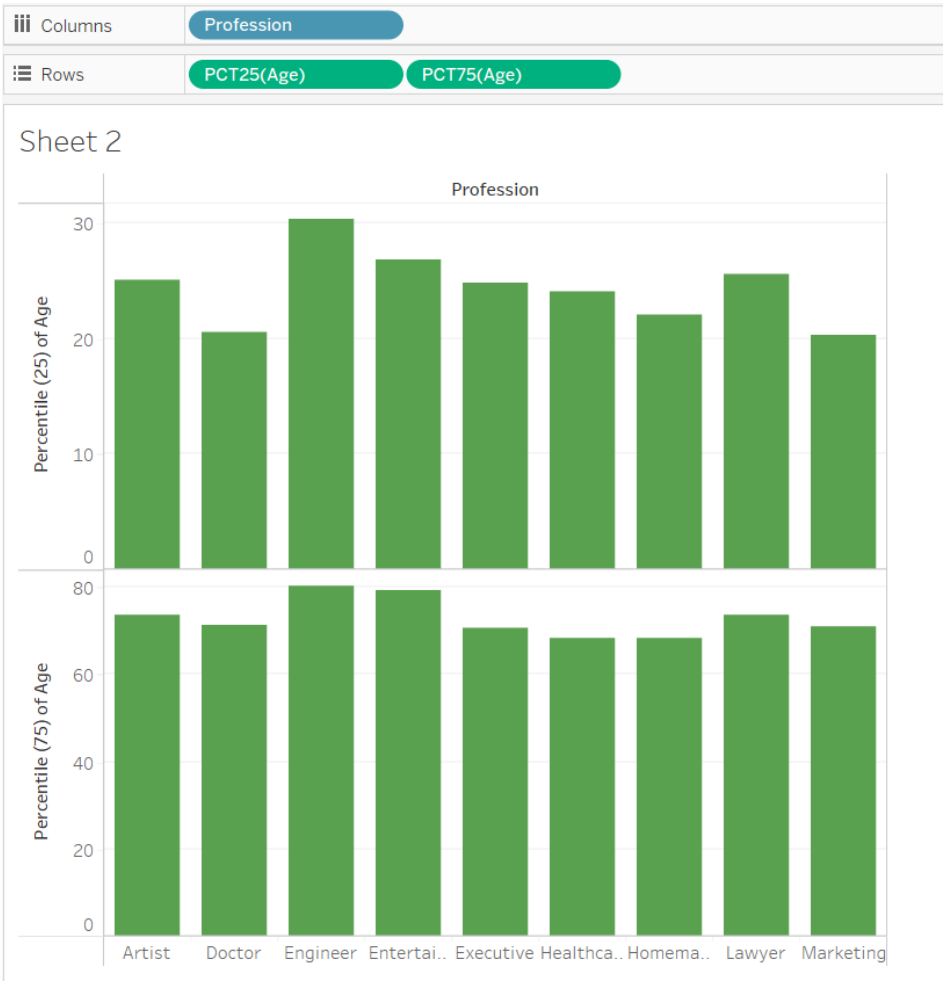
Mean, Median and interquartile of each profession which explain the ages which are more than 25. Minimum and Maximum of each profession are explained.



Using tableau, we tried to visualize the data for Profession and Average(Age) to match the data which we generated using python code and libraries.



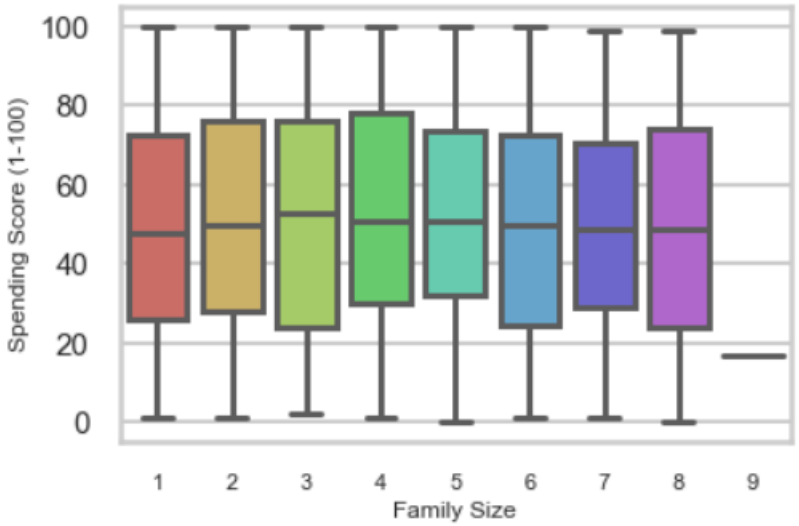
Interquartile of ages verse professions.



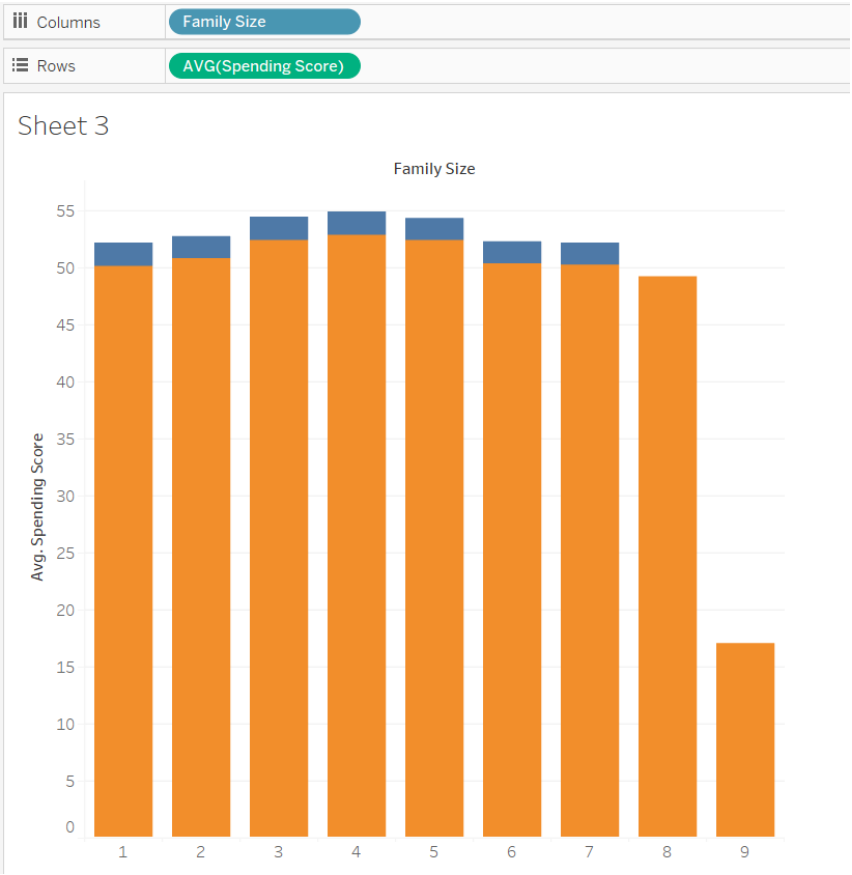
Can we predict the likelihood of a customer making a purchase based on their Family Size and Spending Score.

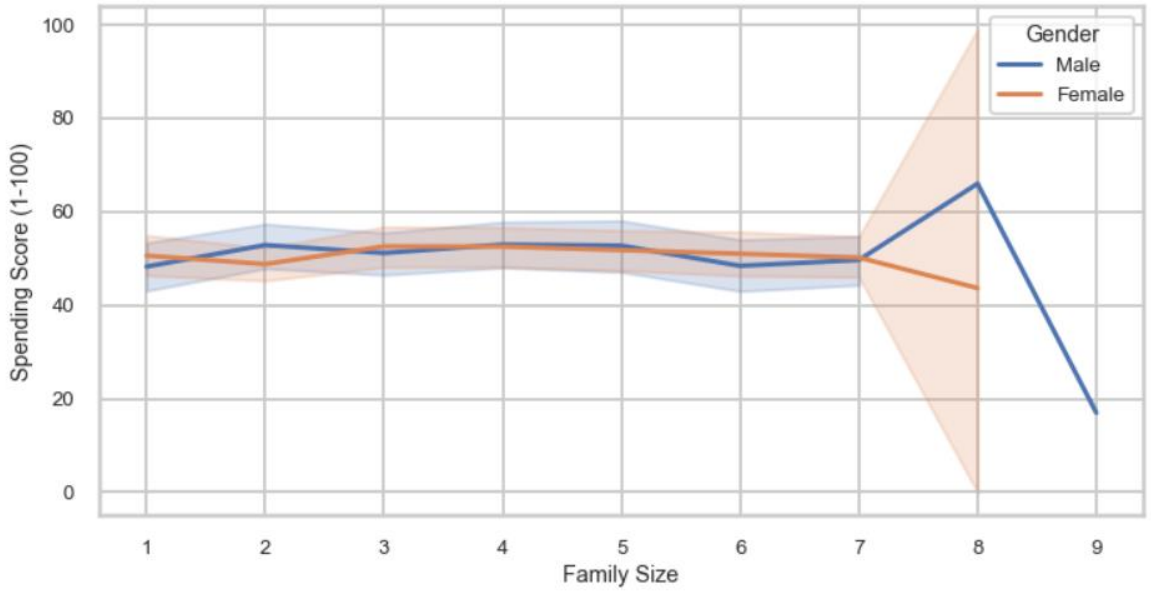
Using the Bar plot, tried to figure out the Mean, max of the spending score verse family size.

In [90]:
```
df_Family_Size  = df['Family Size'].unique()

for Size in df_Family_Size:
    mean_size_value = df.loc[df['Family Size'] == Size, 'Spending Score (1-100)'].mean()
    median_size_value = df.loc[df['Family Size'] == Size, 'Spending Score (1-100)'].median()
    mode_size_value = df.loc[df['Family Size'] == Size, 'Spending Score (1-100)'].mode()
    print(f"Mean score for Family Size {Size}: {mean_size_value:.2f}")
    print(f"Median score for Family Size {Size}: {median_size_value:.2f}")
    print(f"Mode score for Family Size {Size}: {mode_size_value.values[0]:.2f}")
```



From the plot the individual family size are having the mean, median and Interquartile.



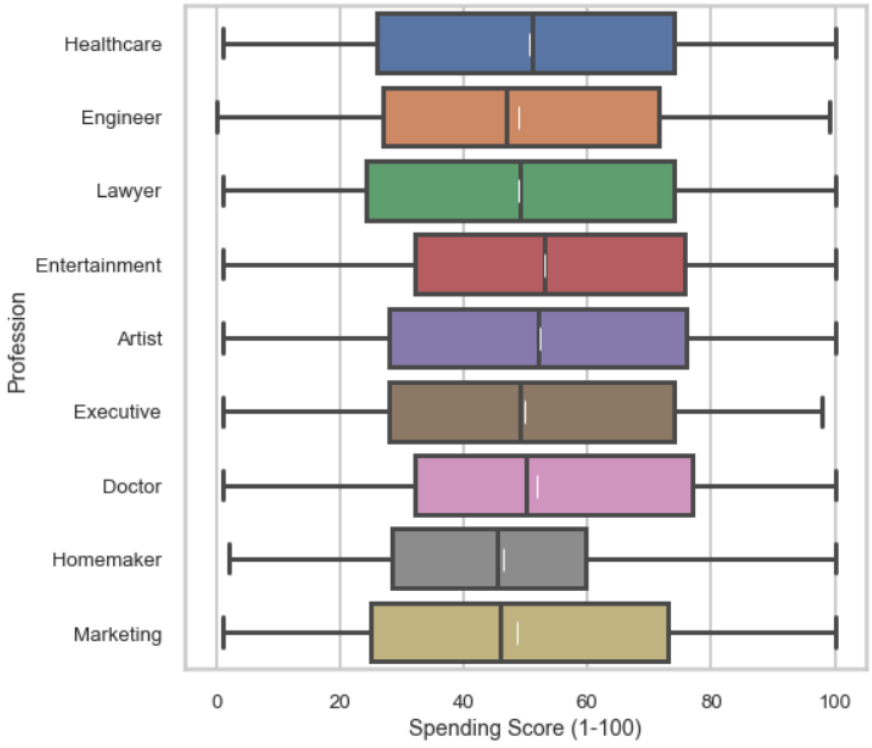


The above line plot explains about how the spending scores are varying with respect to Family size.

What is the likelihood of a customer having a high spending score based on their profession?

```
df_professions = df['Profession'].unique()
#mean_age = []

for profession in df_professions:
    mean_score_value = df.loc[df['Profession'] == profession, 'Spending Score (1-100)'].mean()
    median_score_value = df.loc[df['Profession'] == profession, 'Spending Score (1-100)'].median()
    mode_score_value = df.loc[df['Profession'] == profession, 'Spending Score (1-100)'].mode()
    print(f"Mean Spending Score (1-100) for profession {profession}: {mean_score_value:.2f}")
    print(f"Median Spending Score (1-100) for profession {profession}: {median_score_value:.2f}")
    print(f"Mode Spending Score (1-100) for profession {profession}: {mode_score_value.values[0]:.2f}")
```

Mean, Median, Mode of Profession with respect to spending score (1-100). Plot explains the mean, median and interquartile.

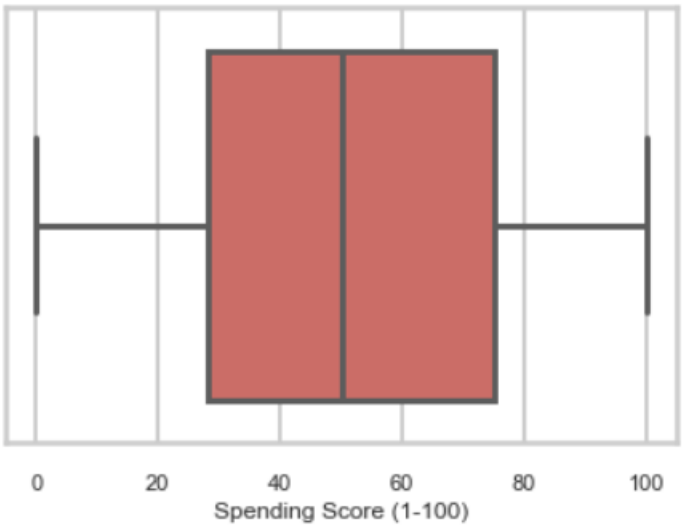

Healthcare
Engineer
Lawyer
Entertainment
Artist
Executive
Doctor
Homemaker
Marketing
Profession
0
20
40
60
80
100
Spending Score (1-100)

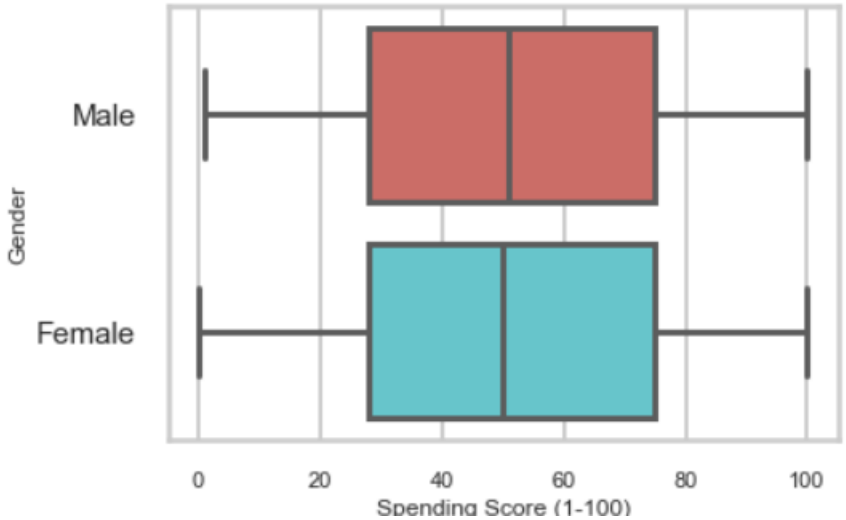

Columns
Profession
Rows
AVG(Spending Score)
Sheet 3
Profession
55
50
45
40
35
30
25
20
15
10
5
0
Avg. Spending Score
Artist
Doctor
Engineer
Entertai..
Executive
Healthca..
Homema..
Lawyer
Marketing

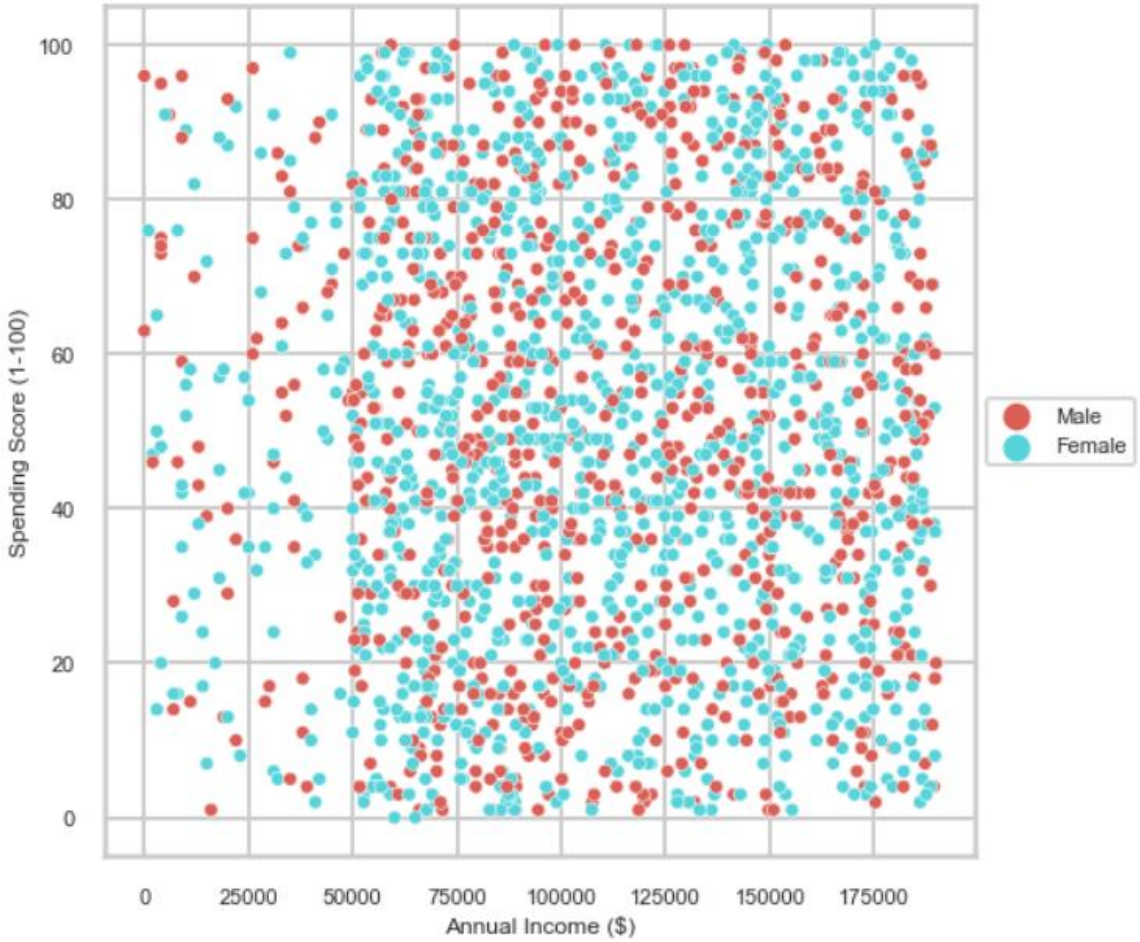
How does the annual Income of customers relate to their spending score, particularly for those with the highest spending score? Can we use their relationship to predict the possible range of annual income for a customer with the highest spending score?

```
Male_df = df[df['Gender'] == 'Male']
Female_df = df[df['Gender'] == 'Female']
#mean of Spending Score
Male_mean_Spending_Score = Male_df['Spending Score (1-100)'].mean()
Female_mean_Spending_Score = Female_df['Spending Score (1-100)'].mean()
print("Male mean of Spending Score (1-100) : ",Male_mean_Spending_Score)
print("Female mean of Spending Score (1-100) : ",Female_mean_Spending_Score)
#median of Spending Score
Male_median_Spending_Score = Male_df['Spending Score (1-100)'].median()
Female_median_Spending_Score = Female_df['Spending Score (1-100)'].median()
print("Male median of Spending Score (1-100) : ",Male_median_Spending_Score)
print("Female median of Spending Score (1-100) : ",Female_median_Spending_Score)
#mode of Spending Score
Male_mode_Spending_Score = Male_df['Spending Score (1-100)'].mode()
Female_mode_Spending_Score = Female_df['Spending Score (1-100)'].mode()
print("Male mode of Spending Score (1-100) : ",Male_mode_Spending_Score)
print("Female mode of Spending Score (1-100) : ",Female_mode_Spending_Score)
```

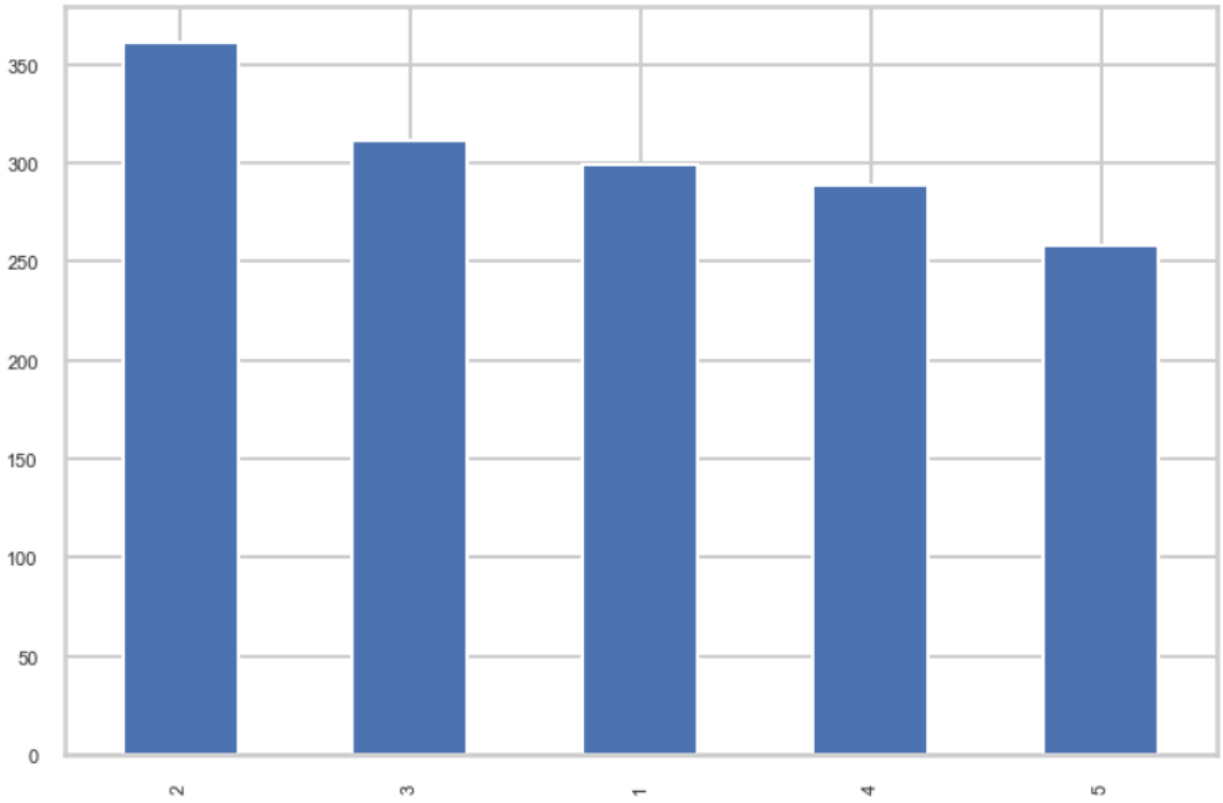
The mean, Median and Interquartile range are statistical measures that describe the distribution of spending scores relative to the annual income of customers.




Male
Female
Gender
0
20
40
60
80
100
Spending Score (1-100)


100
80
60
40
20
0
Spending Score (1-100)
0
25000
50000
75000
100000
125000
150000
175000
Annual Income ($)
Male
Female

What Family Size group has the highest frequency of shopping among customers?



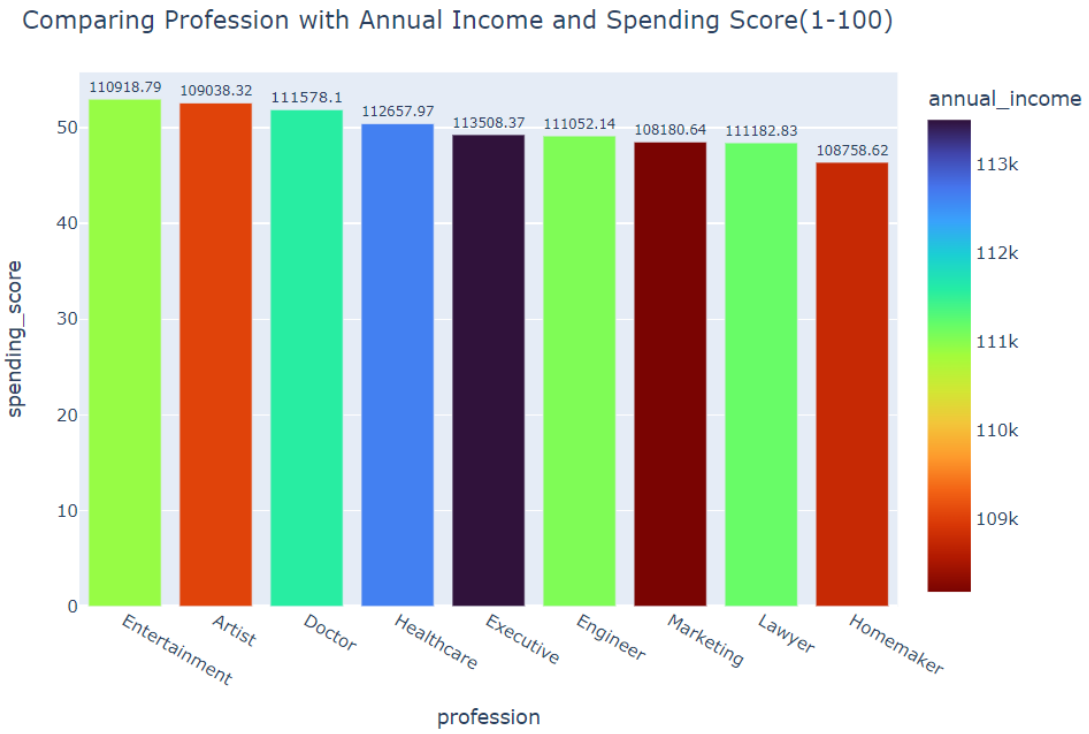
It has been observed that people who frequently shop have a higher likelihood of belonging to households with 2, 3, or 1 family members in that order, compared to households with 4 or 5 family members.

What is the most common profession that has both a high mean income and a high mean spending score among customers?

In [120]:
```
################################Question 7#####################################
#what profession are having mean income and mean spending score?
prof_mean_inc = round(df.groupby("Profession")[["Annual Income ($)", "Spending Score (1-100)"]].mean(),2).reset_index()
prof_mean_inc.columns = ["profession", "annual_income", "spending_score"]
prof_mean_inc = prof_mean_inc.sort_values("spending_score", ascending = False)
prof_mean_inc
```

Out[120]:

	profession	annual_income	spending_score
3	Entertainment	110918.79	52.98
0	Artist	109038.32	52.58
1	Doctor	111578.10	51.88
5	Healthcare	112657.97	50.42
4	Executive	113508.37	49.27
2	Engineer	111052.14	49.14
8	Marketing	108180.64	48.51
7	Lawyer	111182.83	48.42
6	Homemaker	108758.62	46.38


Comparing Profession with Annual Income and Spending Score(1-100)
spending_score
profession
annual_income
113k
112k
111k
110k
109k
0
10
20
30
40
50
110918.79
109038.32
111578.1
112657.97
113508.37
111052.14
108180.64
111182.83
108758.62
Entertainment
Artist
Doctor
Healthcare
Executive
Engineer
Marketing
Lawyer
Homemaker